Case Studies on Transport Policy 12 (2023) 100933

Contents lists available at ScienceDirect

Case Studies on Transport Policy

journal homepage: www.elsevier.com/locate/cstp

Investigating the effects of POI-based land use on traffic accidents in Suzhou Industrial Park, China

Hyungchul Chung [a], Qiaonan Duan [a], Zihao Chen [b,*], Yueming Yang [a]

[a] Department of Urban Planning and Design, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[b] Department of Mathematics and Statistics, University of Exeter, Exeter, UK

ARTICLE INFO

Keywords:
Point-Of-Interest (POI)
Land use
Built environment
Spatial variation
Accidents
Geographically Weighted Negative Binomial Regression (GWNBR)

ABSTRACT

The development of city transportation often promotes prosperity in urban economies, but it can also cause safety issues that adversely affect local residents. It is widely acknowledged that the built environment has a significant impact on traffic accidents. Many studies have shown a link between different land use characteristics and accidents, although these investigations have used different land use characteristic variables, and but few have attempted to allow for spatial heterogeneity. To better understand how urban land use affects the frequency of traffic accidents, this study undertakes a quantitative analysis of the spatial relationship by analyzing the case study of Suzhou Industrial Park, China. We introduce a term frequency-inverse document frequency (TF-IDF) algorithm to identify land use characteristics based on point of interest (POI) data and adopt Geographically Weighted Negative Binomial Regression (GWNBR) model to accomplish this task. This study confirms that local models outperform global models in terms of their explanatory power. The modeling results demonstrate different effects of POI-based land use on traffic accidents and verify the assumption that inherent non-stationarity exists in the estimation of parameters across space.

1. Introduction

With the accelerated pace of urbanization in China, soaring population growth in urban areas has posed unprecedented pressure on the sustainability of cities, especially in terms of traffic issues (Lyu et al., 2020). The development of transportation often promotes prosperity in urban economies, but this transportation development can also cause safety problems that significantly impact local residents. Traffic accidents are now one of the leading causes of death worldwide. Globally, nearly 1.3 million people die in traffic accidents each year, 93 % of which occur in low- and middle-income countries (WHO, 2021). In China, traffic accidents have become the third leading cause of premature death (Xie and Nan, 2018). With the rapid growth of car ownership in China, urban traffic continues to grow. According to China's Traffic Management Bureau (2018), the number of motor vehicles in China has reached 322 million, with 403 million drivers. Traffic safety risks continue to rise and cannot be ignored.

It is commonly believed that the built environment has a considerable impact on the emergence of traffic accidents (Osama and Sayed, 2017; Saha et al., 2020; Li et al., 2013; Yang et al., 2021). The built environment consists of urban design, land use, and transportation systems, and is where the patterns of human activities are highly correlated with the physical environment (Handy et al., 2002). As a significant built environment factor, land use not only affects the demographic and socioeconomic characteristics of a municipality, but also determines traffic volume and traffic behavior patterns (Pulugurtha et al., 2013). Traffic volume has always been a significant predictor of accidents (Dumbaugh and Li, 2010). At present, land development in China is undergoing a change from increment to inventory, and land-intensive development is dominating the regeneration and transformation of old urban areas, which are characterized by high-intensity and high-density development (Zheng et al., 2020). The high-intensity development of urban land in China has led to population agglomeration and traffic congestion, causing serious road safety problems. Road safety can be improved by identifying the factors that mediate the frequency of accidents (Saha et al., 2020). These factors include the planning and policy decisions that feed into the design of the built environment, such as designating specific land uses and developments along urban streets (Johansson, 2009; Ewing and Cervero, 2010). An in-depth understanding of how land use characteristics affect accident risk can help policy makers and transportation planners formulate effective strategies to improve traffic safety.

* Corresponding author.
E-mail address: Zihao.Chen@exeter.ac.uk (Z. Chen).

https://doi.org/10.1016/j.cstp.2022.100933
Received 20 January 2022; Received in revised form 24 September 2022; Accepted 30 November 2022
Available online 5 December 2022

A considerable number of studies have attempted to investigate the relationship between land use and traffic accidents (Peera et al., 2019; Kim et al., 2006; Kim et al., 2010; Ukkusuri et al., 2011; Pulugurtha et al., 2013). Some researchers observed that residential areas are positively correlated with traffic accidents, whereas mixed land use is negatively correlated (Peera et al., 2019). Land use activities that generate traffic volume, such as commercial areas, retail centers, and public parks, have been found to have positive correlations with traffic accidents (Kim et al., 2006; Kim et al., 2010; Ukkusuri et al., 2011; Pulugurtha et al., 2013). Although many studies have established a link between different land use characteristics and accidents, few have taken into account spatial nonstationarity and heterogeneity of such relationship.

Accurate acquisition of land use characteristics is very important in studies of traffic analysis. Commonly used methods including the conversion of high-resolution remote sensing images and land use mapping techniques have problems in identification and accuracy (Zhang et al., 2017; Yao et al., 2017). Point of Interest (POI) is an emerging type of geospatial data widely used at present, offering distinct advantages in that it has larger sample sizes and provides more complete and accurate geographic information, as well as possibly being more up-to-date (Tu and Chen, 2020; Kwan, 2004; Becker et al., 2011; Gao and Zhang, 2018). Land use identification based on POI data not only identifies land use types of the specific unit, but it also describes land use mix level using entropy index, so as to capture more accurate land use characteristics. Compared to traditional data (mapping techniques), POI presents a more detailed picture of multiple land uses (Louw and Bruinsma, 2006). We conclude that POI data is more effective for describing the land use characteristics of a local area than traditional methods.

The majority of the existing studies have relied on global statistical models to analyze the relationship between traffic accidents and various characteristics of the built environment (Huang et al., 2018). However, global models assume that the relationships between variables are consistent over space. Yet, as Lyu et al. (2020) have stated, it is unreasonable to expect that one factor has uniform impact on the dependent variable in different locations across the study space. Neglecting local variations might lead to reduced model reliability and the misinterpretation of the relationship between variables (Qian and Ukkusuri, 2015). Therefore, local models such as the Geographically Weighted Regression (GWR) model (Brunsdon et al., 1996), is gaining popularity. Among the various spatial models in use, GWR is unique in that it estimates local parameters for a regression model by allowing the relationship between the dependent variable and independent variables to vary over space (Huang et al., 2018). The inclusion of spatial nonstationarity and heterogeneity using local models provides a better platform, allowing researchers to explore the the spatial variation of the relationship between traffic accidents and land use factors. Yet, Yue et al. (2017) pointed out that GWR assumes that all explanatory variables share the same space scale; however, the built environment has different effects on human activities in different locations of a city. Based on this, Multiscale Geographically Weighted Regression (MGWR), which allows the scale for the coefficients of each covariate to vary, can be used to capture more accurate spatial relationships (Fotheringham et al., 2017). More recently, Da Silva and Rodrigues (2014) further extend the traditional GWR model family by proposing the Geographically Weighted Negative Binomial Regression (GWNBR) model, which takes into account the over-dispersion of some discrete count data.

The primary objective of this study is to first identify land use characteristics in POI data, then investigate the spatial effects of POI-based land use on traffic accidents at the district level. Through this study, we aim to answer the following questions: (1) Can POI data be used to identify land use characteristics, and how can this be accomplished; (2) Is the GWNBR model useful in capturing spatial nonstationarity in order to investigate the effects of POI-based land use on traffic accidents; (3) Can the GWNBR model outperform a traditional linear regression model in identifying the relationship between traffic accidents and their risk factors. This study enriches existing literature in two aspects. Firstly, we use POI data to identify land use characteristics, which we then apply to a correlation analysis between land use and traffic accidents. Secondly, our study challenges the common assumption, held by global models, that land use factors impact traffic accidents in the same manner across the entire study space. This study is designed to test our hypothesis that GWNBR allows adequate local spatial variation of effects of the same factor on traffic accidents, whereas using a global regression model falls short of doing the same. To test our hypothesis and evaluate the modeling results, we conduct a case study in Suzhou Industrial Park, China. The variations of Suzhou land use factors over space are verified and visualized, which provides valuable guidance to policy makers and transportation planners.

This paper is organized as follows. Section 2 reviews existing studies that extracted significant factors of the built environment that impacts the emegence of traffic accidents, including land use characteristics, road conditions, and accessibility to facilities. Section 3 introduces the study area of Suzhou, China and the data source. Section 4 presents the identification method, process and result of POI-based land use in Suzhou, where land use types and mixed land use are identified respectively. Section 5 elaborates the formulation and estimation of the regression models used, including Negative Binomial Regression (NB), and Geographically Weighted Negative Binomial Regression (GWNBR). Section 6 concludes the paper with an overview of major findings, discussion and potential areas for future work.

2. Literature review

2.1. Influencing factors of traffic accidents

This section will examine individual studies and their findings in detail in order to identify the predicting variables of traffic accident for the case study of Suzhou. Existing research speculating the influencing factors of traffic accidents can be organized into three categories: the natural environment, the built environment, and socio-demographic variables. Below are some examples of factors in these categories.

Characteristics of the natural environment can be divided into climatic conditions and topographic conditions (Xie and Nan, 2018). Anarkooli et al. (2017) found that climate-related factors, such as rainy weather, significantly affect the severity of injury caused by rollover crashes. Xia et al. (2014) discovered that precipitation, temperature, and visibility are strongly correlated with traffic accidents. Topographic factors, such as slope, also have impact on the emergence of traffic accidents (Ukkusuri et al., 2012).

Of all the socio-demographic factors, population is generally found to highly influence the occurrence of traffic accidents (Pulugurtha et al., 2013; Dumbaugh and Li, 2010; Zheng et al., 2020; Osama and Sayed, 2017). An analysis performed by Aguero-Valverde and Jovanis (2006) suggests that the number of low-income households in an area combined with their age statistics impacts the occurrence of traffic accidents. Regional employment density and education levels of households are also verified by Pulugurtha et al. (2013) and Dumbaugh and Li (2010) to be influencing factors. Moreover, a study by Cheng and Liu (2017) found that migration into/out of an area has the potential to influence the occurrence rate of traffic accidents.

Built environment has been the prime interest when considering the influence of spatial features in modelling (Lyu et al., 2020). The introduction section above briefly mentioned the signs of the built environment having a significant impact on traffic accidents. Built environment factors in this study are measured using the “six Ds” summarized by Cervero and Kockelman (1997) and Ewing and Cervero (2010), including density, diversity, design, destination accessibility, distance to transit, demand management. It should also be noted that the above variables are roughly categorized without determined boundaries yet (Ewing and Cervero, 2010). Further works documenting the relationship between the built environment and accidents are classified and listed in

Table 2.1.

Density is commonly measured as the variable of interest per unit of area (Ewing and Cervero, 2010). The "variable of interest" can be population, employment, land use type, etc. Almost every study reviewed in this paper takes population density into consideration, although it falls into the category of demographics. It was consistently found that there was a significant positive correlation between population density and the occurrence of traffic accidents (Casares et al., 2019; Saha et al., 2020; Wier et al., 2009). In contrast, the effects of land use on accidents varied. Most studies revealed that the presence of commercial land was correlated with a higher concentration of road traffic accidents (Osama and Sayed, 2017; Huang et al., 2018; Dumbaugh and Li, 2010). Residential land use was found to have distinct effects on traffic accidents in different scenarios, with negative correlations between them found in some studies (Ukkusuri et al., 2012; Osama and Sayed, 2017) and positive ones found in the others (Pulugurtha et al., 2013; Delmelle et al., 2012). A divergence of effect also appeared regarding industrial land (Ukkusuri et al. 2012; Chen and Zhou, 2016). Ukkusuri et al. (2012) demonstrated that industrial land had a positive impact on accident occurrence, possibly due to higher speed limits; on the other hand, Chen and Zhou (2016) found that that the efficiency of traffic flow in industrial areas decreased accident rate in the areas accordingly. In addition, Osama and Sayed (2017) discovered a decline in crashes associated with the increase of recreational areas. Many reasons could account for the inconsistent findings, hence the importance of continued study of the association between land use and traffic accidents.

Diversity in the context of built environment refers to the number of different types of land use within a given area, as well as the extent to which they represent a geographic area. It has been included in model building in many studies (Ewing and Cervero, 2010; Wier et al., 2009; Pulugurtha et al., 2013; Ding et al., 2018; Peera et al., 2019; Casares et al., 2019). The diversity of land use – in other words, mixed land use – has been measured in various ways and has been found to have different effects on the frequency of the occurrence of traffic accidents. Peera et al. (2019) found that mixed land use assessed using a polygon shapefile was shown to be negatively correlated with road traffic accidents, wheareas mixed land use in terms of the entropy index was found to have a positive but limited effect on road safety (Casares et al., 2019). However, in the study conducted by Wier et al. (2009), the prediction model indicated that a higher proportion of mixed residential/commercial land use led to an increase in vehicle–pedestrian injury collisions. Moreover, density and mixed land use measured by the entropy index were shown to have major effects on the number of crashes, accounting for approximately 66 % of crashes in total in one particular area (Ding et al., 2018). From these studies, it is clear that an effective way to measure mixed land use is needed for our research.

One aspect of design includes road network characteristics within an area. Research into road design has mainly focused on road lengths, road curvature, intersections, signals, speed limit settings etc. (Osama and Sayed, 2017; Ding et al., 2018; Huang et al., 2018; Saha et al., 2020, Tercan et al., 2020, Patil and Pawar, 2016). There is a significant amount of evidence demonstrating that longer road lengths or higher proportions of arterial road could lead to a greater concentration of road accidents (Casares et al., 2019; Osama and Sayed, 2017; Peera et al., 2019; Dumbaugh and Li, 2010). Investigators have also observed that a higher density of four-way intersections in a block group usually results in a higher crash rate (Huang et al., 2018; Saha et al., 2020). Patil and Pawar (2016) highlighted the role that speed plays in the emergence of traffic accidents, with a positive correlation found between accident rate and speed differential in traffic stream. Osama and Sayed (2017) also analyzed traffic signal and light pole densities, both of which were shown to positively impact the number of crashes.

Destination accessibility refers to ease of access to trip attractions (Ewing and Cervero, 2010). In this context, destination accessibility corresponds with the number of schools, parks, or other attractions reachable within a study unit. Typical measures used in prior studies of destination accessibility include numbers of schools and recreational

Table 2.1
Summary of literature review.

D variables	built environment factors	Wier et al., 2009	Dumbaugh & Li, 2010	Ukkusuri et al., 2012	Pulugurtha et al., 2013	Dai & Jaworski, 2016	Osama & Sayed, 2017	Huang et al., 2018	Ding, Chen, & Jiao, 2018	Casares et al., 2019	Saha et al., 2020
Demographics	population	√	√	√	–	√	√	√	√	√	√
Density	commercial land	√	√	√	√	–	√	√	√	–	√
	residential land	√	–	√	√	–	√	–	–	–	–
	industrial land	√	–	√	√	–	–	–	√	–	√
	recreational park	–	–	–	√	–	√	–	√	–	–
	offices	–	–	√	√	–	–	–	–	–	√
	retails	–	√	√	–	–	–	–	–	–	–
	governmental land	–	–	–	√	–	–	√	√	–	–
Diversity	land mixed use	√	–	–	√	–	–	√	√	√	–
Design	road	√	√	√	√	√	√	√	√	√	√
	intersections	√	√	√	–	√	–	√	–	–	√
	signals	–	–	–	–	–	√	–	√	–	–
	light poles	–	–	–	–	–	√	–	–	–	–
Destination Accessibility	number of restautants	–	–	–	–	√	–	–	–	–	–
	number of schools	–	–	√	–	–	–	–	√	–	√
	number of parks	–	–	√	–	–	–	–	–	√	√
	number of hotels	–	–	–	–	–	–	–	–	√	–
Distance to Transit	number of bus stops	–	–	√	–	√	√	–	√	–	√
	number of railway stations	–	–	√	–	√	√	–	–	–	–
Demand Management	parking space	–	–	√	–	–	–	–	–	–	–

parks. In a study performed by Ukkusuri et al. (2012), researchers found that study units containing a greater number of schools are more likely to have more traffic accidents. Casares et al. (2019) analyzed tourist-oriented buildings and environments in relation to traffic accidents and found that tourism is beneficial for improving road safety in a given area.

Distance to transit is usually measured as the number of transit stations per unit area. Some studies consider bus stops and subway stops as separate variables (Dai and Jaworski, 2016; Osama and Sayed, 2017). Others combine them into a single variable or consider only bus stops (Ding et al., 2018; Saha et al., 2020). Public transit stops tend to be positively associated with traffic accidents, as confirmed by Osama and Sayed (2017), Dai and Jaworski (2016), among others. A few studies attempt to take parking space into account, which can be considered as demand management of vehicle volume, such as the work of Ukkusuri, et al. (2012). The impact of parking space on accidents, however, is unclear.

2.2. Application prospect of POI data in land use

Currently, land cover information is the main data source used in studies of urban growth simulation (Li et al., 2015; Li et al., 2014). However, land cover information is not easily applicable to studies of urban transportation and land use patterns (Xue et al., 2015). Land cover is a physical description of land surface (Hu et al., 2016), while land use is a cultural concept describing human activities and their use of land. Land cover can be used to infer land use, but the result is not always accurate. Moreover, land use maps are often produced by interpreting aerial photographs, field findings, and auxiliary materials such as statistical data (Wu et al., 2009).

As cities rapidly evolve, land use databases are not consistently updated in real time, resulting in out-of-date maps. In contrast, satellite-based remote sensing has certain advantages in monitoring the dynamics of urban land use (Hu and Wang, 2013; Wu et al., 2007). However, getting high-resolution urban land use maps from satellite images remains a daunting and expensive task. Land use information obtained from medium-resolution satellite images cannot provide sufficient separation between urban functional areas, and it is difficult to extract socio-economic features from these images (Liu et al., 2015; Hu et al., 2016).

POI data is an emerging form of geographic data voluntarily provided by individuals mainly for monitoring users' positions in spatial tracking or geocaching systems (Hu et al., 2016). POI data can link geographic locations to descriptive features or other place-based information (McKenzie et al., 2014). The distinct advantages of POI data being accessible, informative, and up-to-date gained itself popularity among researchers (Yang et al., 2018). For example, Kwan (2004) carried out a spatiotemporal simulation of the density and distribution of residents' activities based on POI data. In the United States, Becker et al. (2011) used the POI data of Morriston, Florida to study migration and urban population change. Furthermore, with POI data, Gao and Zhang (2018) used the shortest distance method to analyze the accessibility of health facilities in Xi'an, China. Plenty of applications suggested that employing POI data could be an effective way to identify and obtain land use characteristics.

Many studies have suggested that the distributions of POIs can reveal the land use features of some regions (Sun, et al., 2016; Pan et al., 2019; Bao et al., 2017; Hu et al., 2016). There are three potential advantages of acquiring land use information through POI data for accident risk factor analysis. First, reliable POI data can be collected from anywhere in the world, which thus provides ample space for detailed accident analysis (Pan et al., 2020). Second, given that POI data are specific data of land use factors with precise locational information, studies show that they are strongly related to traffic accidents at the macro and micro levels (Jia et al., 2018). Third, this kind of data details the distribution of urban infrastructure and has much better statistical granularity, which was proven to have good statistical performance at the traffic analysis zone (TAZ) level (Chen et al., 2019; Chen et al., 2020; Pan et al., 2020). For these reasons, we used POI data to describe the built environment, as opposed to traditional land use data. Additionally, we chose three kinds of diversity indexes, as described by Yue et al. (2017), as diversity measures to quantify the diversity of land use patterns.

2.3. Modeling tools

Most existing studies employed generalized linear regression models to explore the association between traffic accidents and their risk factors (e.g., Wier et al., 2009). Poisson Regression model and Negative Binomial (NB) regression model are also widely used in the detection of traffic accident risk factors (Dumbaugh and Li, 2010; Dai and Jaworski, 2016; Osama and Sayed, 2017; Casares et al., 2019; Peera et al., 2019). Some studies also use non-linear models to explore the interaction between built environment factors and the non-linear influence of some factors, such as the Multiple Additive Poisson Regression Trees developed by Ding et al. (2018).

However, the main disadvantage of the global models is that they cannot examine the influence of spatial non-stationarity and heterogeneity. These need to be considered in the planning and management of the built environment in order to adopt geo-spatial differentiated policies, thus promoting safer roads. The Geographically Weighted Regression (GWR) model is specifically designed for this purpose (Brunsdon et al., 1996). GWR can solve spatial autocorrelation problems and has already been employed to study the impact of built environment on crashes (Huang et al., 2018). More recently, Da Silva and Rodrigues (2014) further extend the traditional GWR model by proposing the Geographically Weighted Negative Binomial Regression (GWNBR) model which takes into account the over-dispersion of discrete count data.

3. Research design

3.1. Study area and data

3.1.1. Study area

The area of study is Suzhou Industrial Park (SIP), located in the eastern part of Suzhou, China. Suzhou is one of the major economic centers in the Yangtze River Delta region, and SIP is classified as both a national economic and technological development zone and a national high-tech industrial development zone. The SIP administrative region covers an area of 278 square kilometers. According to 2019 census data, SIP has a registered population of 0.576 million, a floating population of 0.745 million, making a total permanent population of 1.15 million. SIP is also the largest collaboration project between the governments of China and Singapore, drawing on the successful experience of 15 advanced countries and regions in the areas of development and management. Compared with other development zones in China, SIP covers a wider range of land use types and is cut through and divided by several waterways, which adds to the complexity of analysis (Fig. 3.1).

3.1.2. Data

The dataset used in this study includes two categories: traffic accident data and built environment data. Among them, the latter mainly include POIs (e.g., shopping and leisure places, schools, etc.), population, road network and facilities (e.g., traffic lights and intersections, etc.). The research unit used in this study is the traffic analysis zone (TAZ), and all data are georeferenced. The data for these categories is described in more detail below.

Empirical studies highlighted the effectiveness of using observational data to model the influencing factors of the emergence of traffic accidents (Tercan et al., 2020). For this study, the extracted dataset contains records of a total of just under 60,000 traffic accidents occurred in SIP in 2016. Fig. 3.2 shows the spatial distribution of the recorded

Fig. 3.1. Administrative area map of SIP (Source: Baidu Map).

accidents in the dataset. Each traffic accident is represented by a dark red point on the map. After initial processing, a total of 56,958 valid accident data points are left. The dataset provides information of the location of the accident (via longitude and latitude), the gender and age of the person involved, the weather at the time the accident happened, whether there is any injury caused by the accident, and whether the

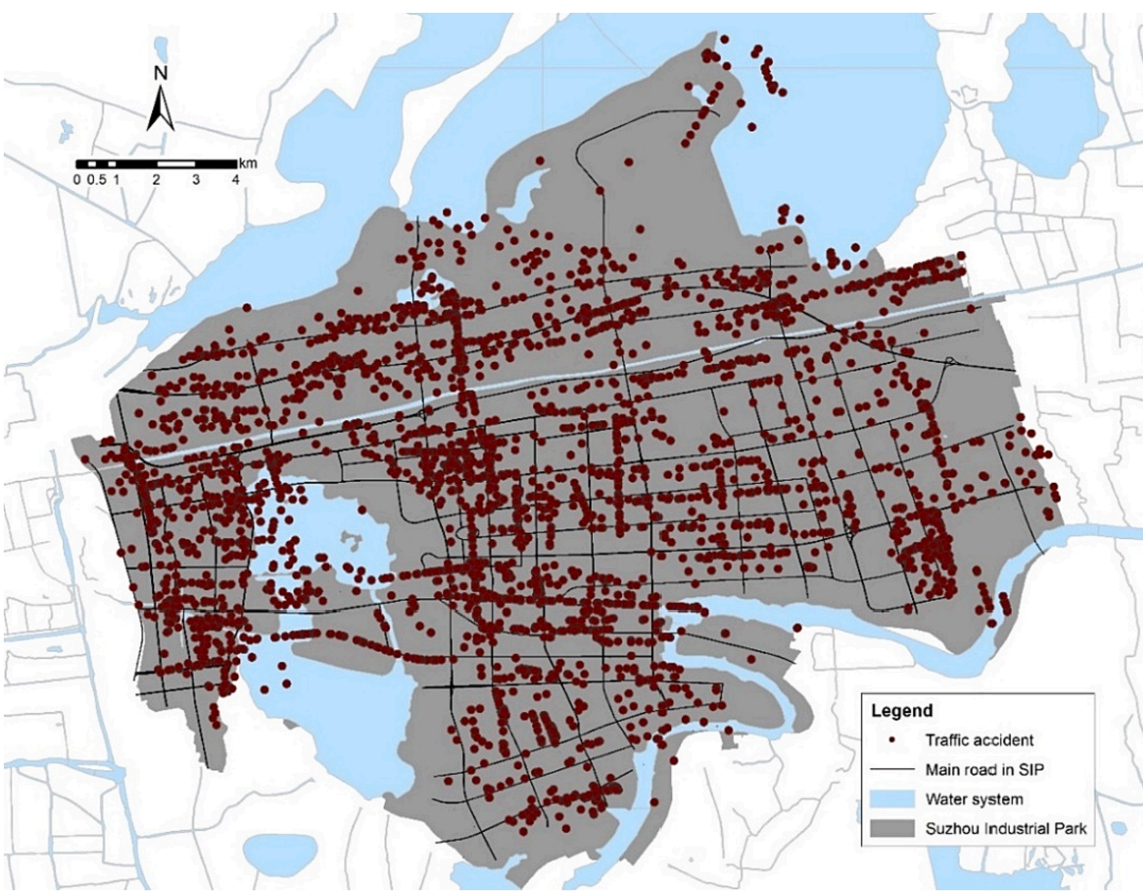

Fig. 3.2. Traffic accident distribution map of SIP in 2016.

accident took place on the road or intersection or junction. Of all the traffic accidents that occurred in SIP in 2016, 52 % ended up with no injury caused, 19 % resulted in wound. 14 % of all the accidents took place on a rainy day. For the purpose of this study, we are only interested in the location aspect of the accidents.

Fig. 3.3 shows the frequency distribution of traffic accidents. Using the Moran's I index, we examine the TAZ spatial autocorrelation of the traffic accidents. The Moran's I value is 0.036 (p-value < 0.01), suggesting that there is a small spatial dependence in accident occurrence (clustering).

On this basis, GeoDA is used to carry out spatial hot spot analysis of accidents, and the clustering area of traffic accidents within SIP is obtained (see Fig. 3.4). According to the cluster map, it is not difficult to find that the high-value clustering occurred in the west and northeast of SIP, while the low-value clustering occurred in the southeast. In addition, combining Figs. 3.3 and 3.4, it can be inferred that the accidents present a relatively discrete distribution, and the number of accidents in adjacent TAZ may vary greatly.

Population data came from the Data Center of The Chinese Academy of Sciences. The dataset reflects the detailed spatial distribution status of population in China. The SIP population data was obtained by refining the grid, after which we conducted kernel density analysis (Fig. 3.5).

Road network and facility data from 2016 was used to ensure consistency and accuracy. Based on previous research (see section 2), we selected the lengths of the roads and the number of intersections as the design variables in this study.

POI data was obtained from the Gaode Map Open Platform (also known as AMap). There are approximately 224,000 POI data points in SIP, which belong to 13 categories including restaurants, accommodation, parks, companies, etc. This data is used to identify the land use characteristics in the next section.

3.2. Operationalization

3.2.1. Identification of POI-based land use

3.2.1.1. Land use type. The raw POI data had many categories, and each category contained multiple levels of smaller categories. There were also duplicates or overlaps between categories, making it necessary to reclassify the original data. For example, dormitories exist not only in residential areas of commercial residences but also in schools that provide scientific, educational, or cultural services. In this context, following Chi et al. (2016), the POI data were divided into seven categories: commercial land, residential land, amenity, public service, industrial land, green land, and transportation. Following the principles of universality and consistency, we present the three levels of POI classification in Table 4.1.

Term frequency-inverse document frequency (TF-IDF) was used to calculate the weights of the different POI categories for each TAZ. This algorithm, proposed by Salton and Buckley (1988), is a statistical method for estimating the importance of a word to a document. The importance of a word increases proportionally with the frequency of its occurrence in the document, but it decreases inversely with the frequency of its occurrence in the corpus (Salton and Buckley, 1988). In other words, the more frequently a word appears in an article and the

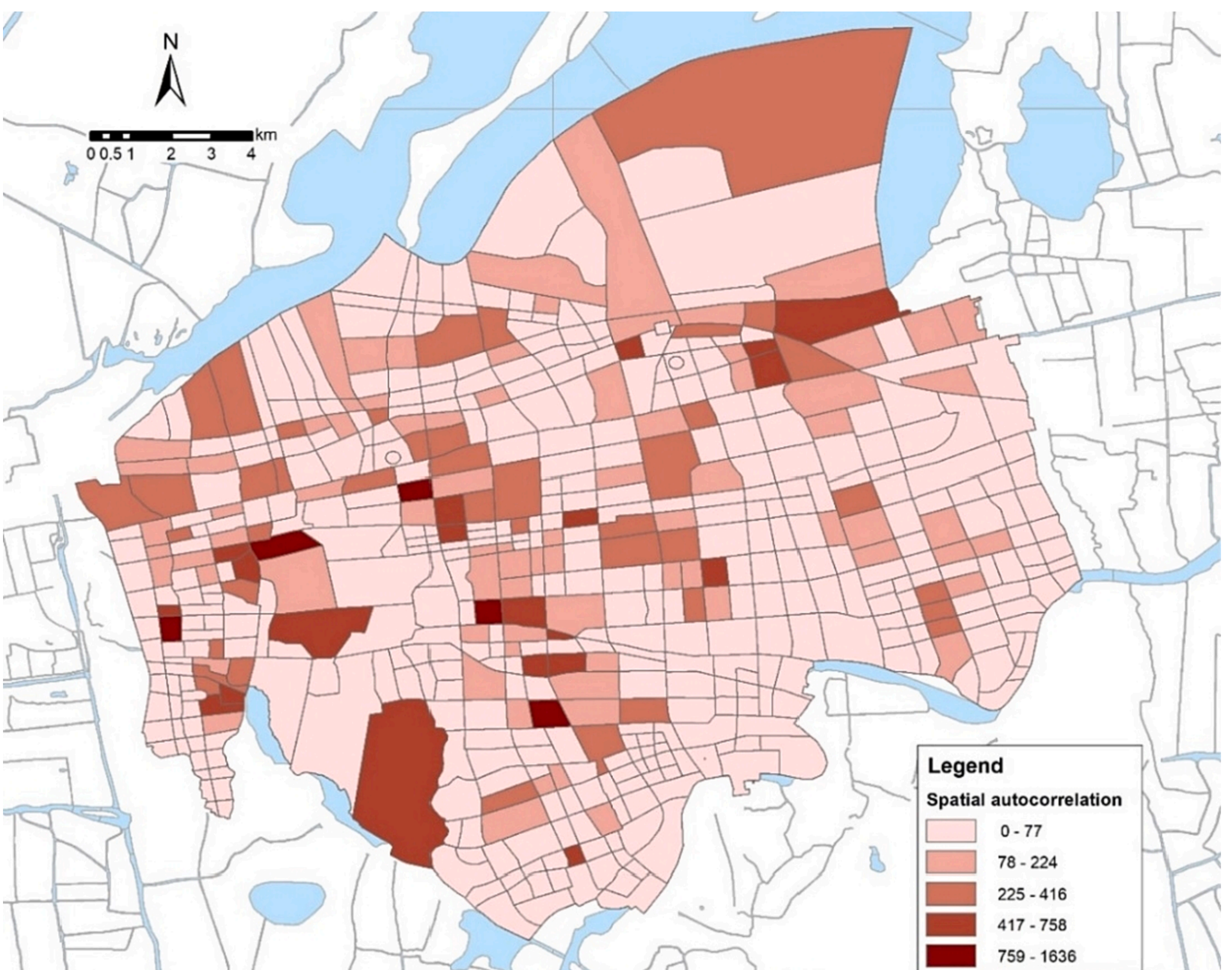

Fig. 3.3. Accident frequency distribution map of SIP in 2016.

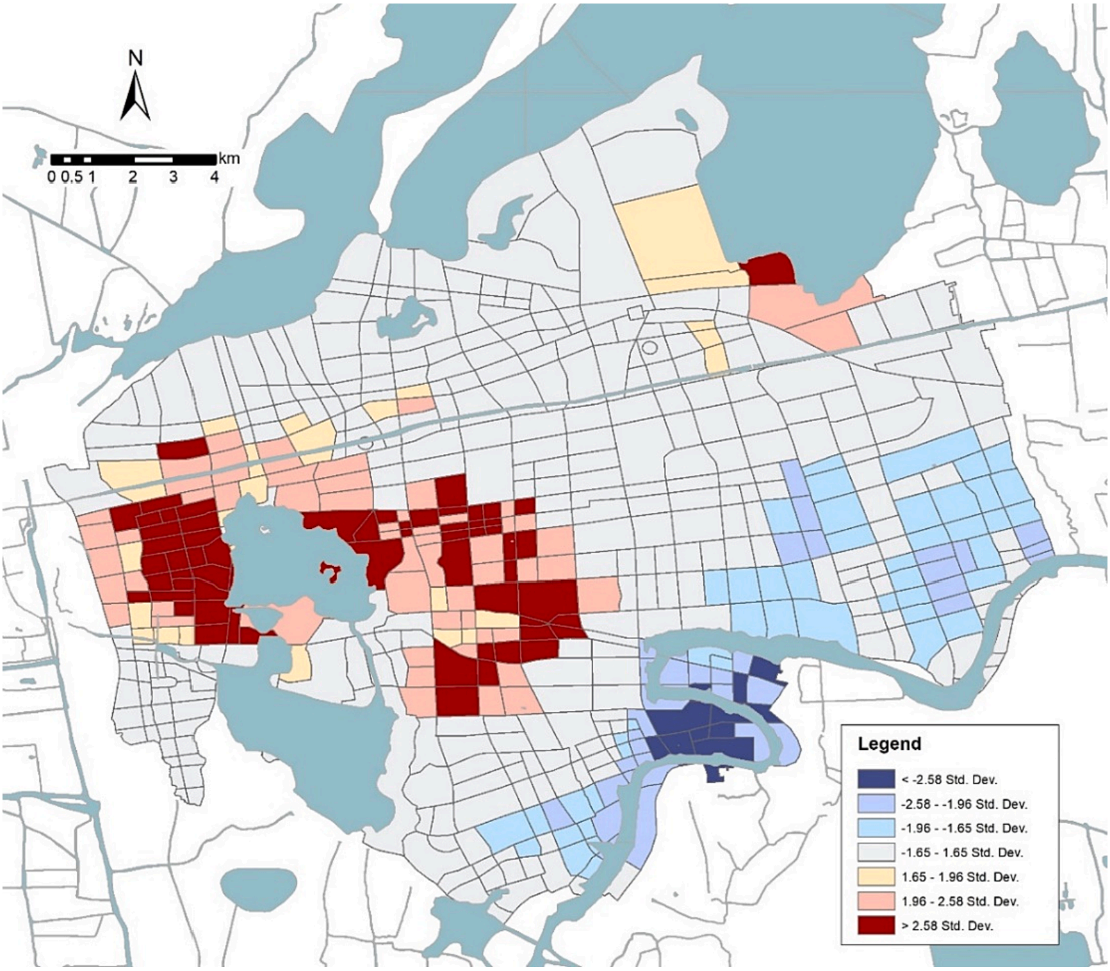

Fig. 3.4. Accident cluster map of SIP in 2016.

less frequently it appears in all documents, the more the word represents the article. TF represents the word frequency, and IDF represents the inverse text frequency index. Due to its strong classification capability, the TF-IDF algorithm is widely used in machine learning (Chen, 2017; Zhao et al., 2019). The calculation formula for TF-IDF is shown in equations (1), (2), and (3).

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

In formula (1), i denotes the word, j stands for the document, $n_{i,j}$ represents the number of times word i appears in document j, and $\sum_k n_{k,j}$ represents the sum of the occurrences of all words in document j.

$$idf_i = log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{2}$$

In formula (2), "$|D|$" stands for the total number of documents, and "$|\{j : t_i \in d_j\}|$" represents the total number of documents containing words.

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{3}$$

The current study regards each spatial unit (TAZ) as a "document" and the categories of the various POIs in the document as "words". Therefore, the analysis of the categories of each TAZ is transformed into solving the weight of each word in each document. The POI category with the highest weight is taken as the main function of the TAZ. The identification process is shown in Fig. 3.6.

Firstly, the preprocessed POI data and TAZs were connected in ArcGIS to obtain the number of the various POIs in each TAZ. Next, the POI categories were merged, and then the name of each POI category and the number of each TAZ were read and written to the document. Secondly, according to the TF-IDF algorithm, the corresponding document of each TAZ was read, and the weighted results of all POI categories were calculated. Finally, the category of the maximum weight in each TAZ was output to obtain the category sequence of all units, after which the land use map was obtained (see Fig. 3.7).

The land use map obtained via POI can also be compared with the official land use map of SIP (Fig. 3.8) for verification purposes. It can be seen that the land use classifications are largely consistent among the two maps.

In research using POI data to identify urban functional areas, base maps are primarily grounded on road networks or grid networks. In our research, we first completed the identification of land use types based on the TAZ unit. The results are shown in Fig. 3.7 above. One finding is that in TAZs with large areas, commercial land accounts for the majority, while there are some residential areas covered as well.

To increase the credibility of the above results, we also used grids for identification (Fig. 3.9). However, too many units divided by the grids caused difficulties for subsequent analysis. Moreover, by comparing Figs. 3.7 and 3.9, it can be seen that the results were similar and that there were no regions with large deviations. In addition, we also compared the results with all land types using remote sensing images and found that the results of the land use-type identification using the TF-IDF algorithm conformed to the current situation of land use in SIP and met the requirements of basic analysis (Table 3.1).

3.2.1.2. Mixed land use. Yue et al. (2017) considered Hill numbers,

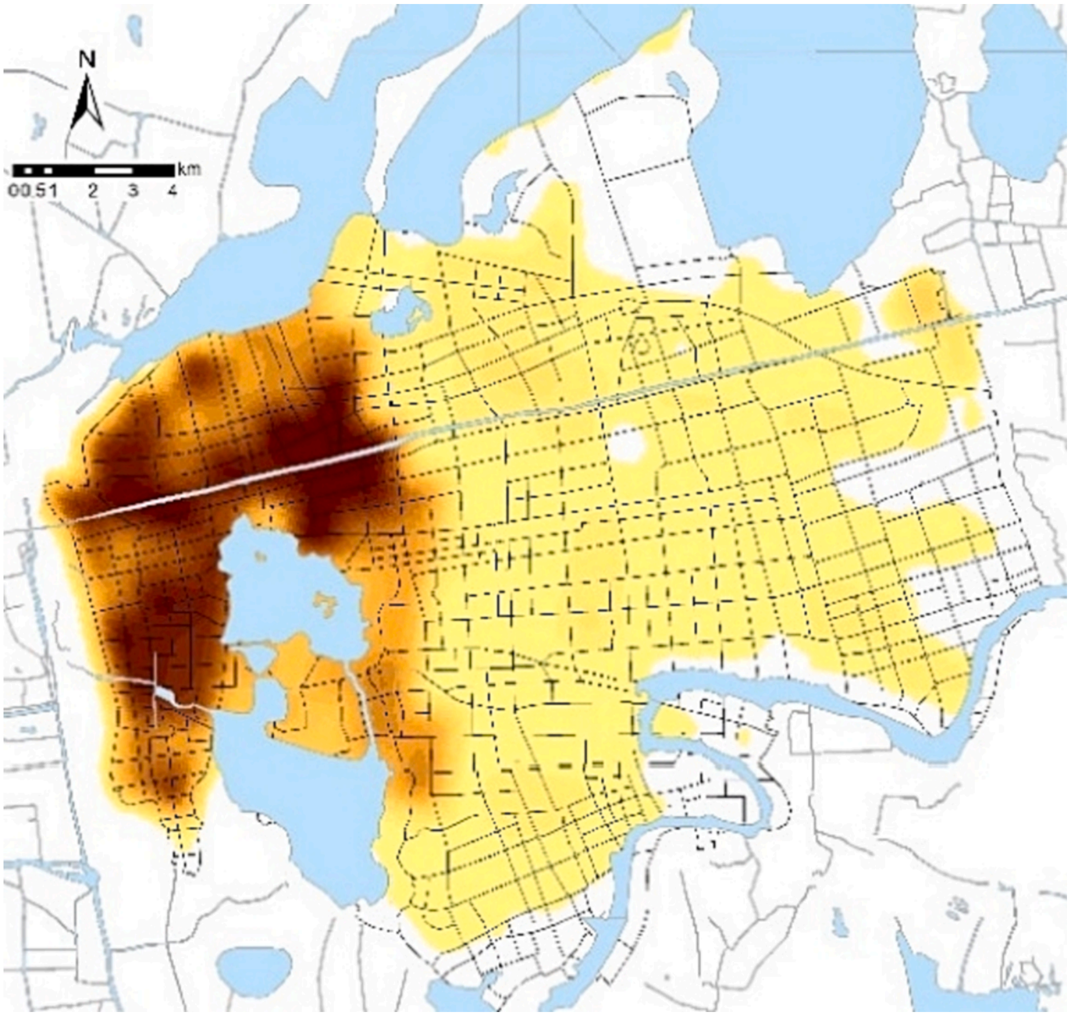

Fig. 3.5. Population distribution of SIP.

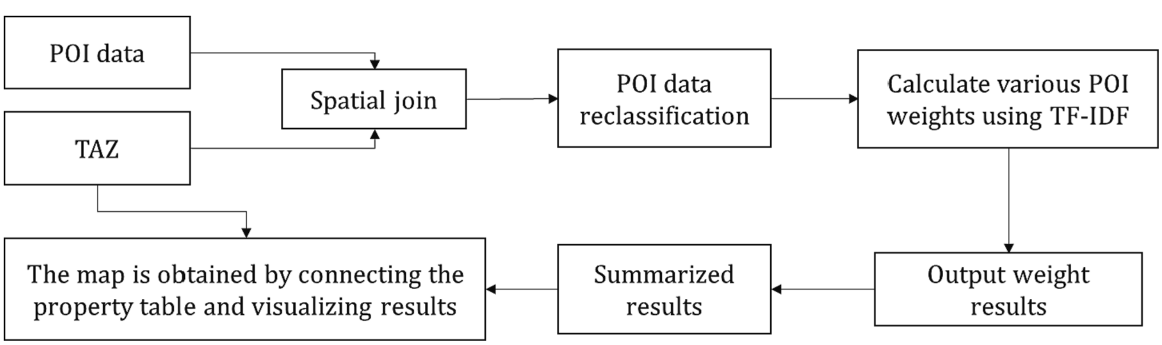

Fig. 3.6. Flow chart of identification of the POI-based land use type.

which is originally developed in the field of ecology and biogeography, as a better measurement of the diversity of POI for reflecting mixed land use. Hill numbers includes richness (Equation (4)), entropy (Equation (5)) and the Simpson index (Equation (6)). In these formulas, ${}^nD$ denotes the diversity index, s denotes the number of species, and P_i denotes the relative abundance of the i-th species.

$$ {}^0\boldsymbol{D} = \sum_{i=1}^{s} \boldsymbol{P}_i^0 \tag{4} $$

In this context, the richness index reports the number of different types of land use present in each TAZ, i.e. the number of POI categories (species). A larger ${}^0D$ value (richer) indicates that there are a greater number of POI categories in a TAZ. However, it does not reflect the number of POIs belonging to a certain category.

$$ {}^1\boldsymbol{D} = \boldsymbol{exp}\left(- \sum_{i=1}^{s} \boldsymbol{P}_i \boldsymbol{ln} \boldsymbol{P}_i \right) \tag{5} $$

How much a piece of information contains is directly related to its uncertainty. According to formula (2), the greater the uncertainty of the variable is, the greater the ${}^1D$ value becomes, and the greater the amount of information is needed to discern it. Shannon entropy has been

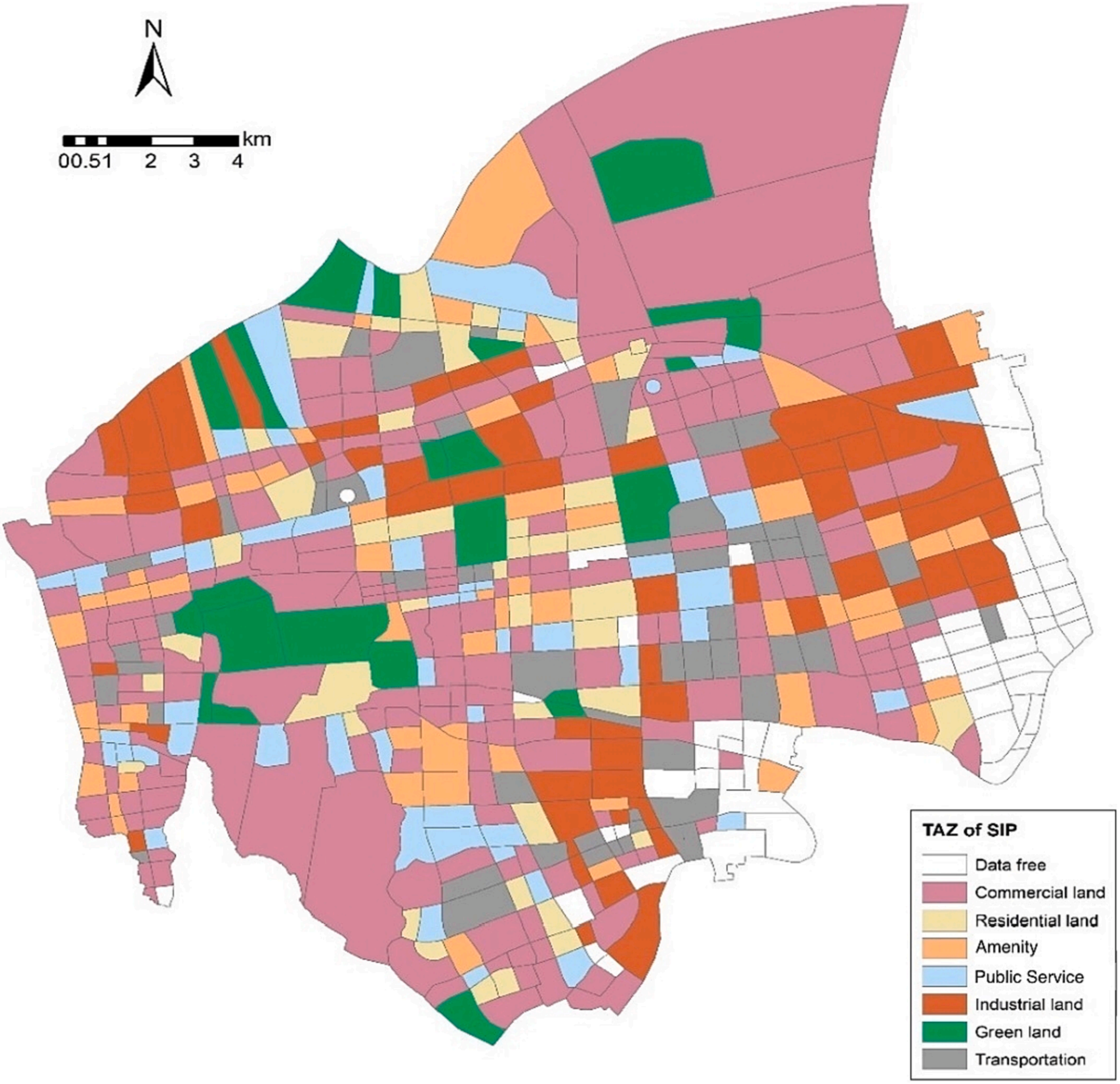

Fig. 3.7. Land use map of SIP based on POI data identification – TAZ.

used widely to measure the diversity of land use, since it reflects the amount of order in both POI categories and in the number of POIs (Casares et al., 2019; Ding et al., 2018; Yue et al., 2017).

$$^{2}\boldsymbol{D} = 1/\left(\sum_{i=1}^{s} \boldsymbol{P}_i^2\right) \tag{6}$$

The Simpson index describes the probability that two randomly selected individuals (POIs) from a sample (TAZ) belongs to the same species (POI category). The larger its value is, the more species there are in the community, the more uniform distribution of various individuals, that is, the higher degree of species diversity.

Richness (POI), entropy (POI), and Simpson (POI) indexes are used to report the mixed land use in SIP based on POI data. The diversity results measured by these indicators are shown in Fig. 3.10. In general, the spatial variation in mixed land use characterized by the three indices tend to be consistent.

3.2.2. Descriptive statistics

The dependent variable is the number of traffic accidents in each TAZ, which is a discrete count variable. The selection of independent variable is based on the seven D variables discussed in section 2. Initially, a total of 19 variables were included in this study. Correlation test showed strong correlation between some independent variables, and therefore some variables were dropped and 12 remaining variables (as shown in Table 3.2) were added to the model. Because of the correlation between the three indicators measuring mixed land use, three different models were constructed, each containing one of the three indexes.

3.3. Methodology

This section introduces the two models used in this study, namely the Negative Binomial (NB) regression, and the Geographically Weighted Negative Binomial Regression (GWNBR).

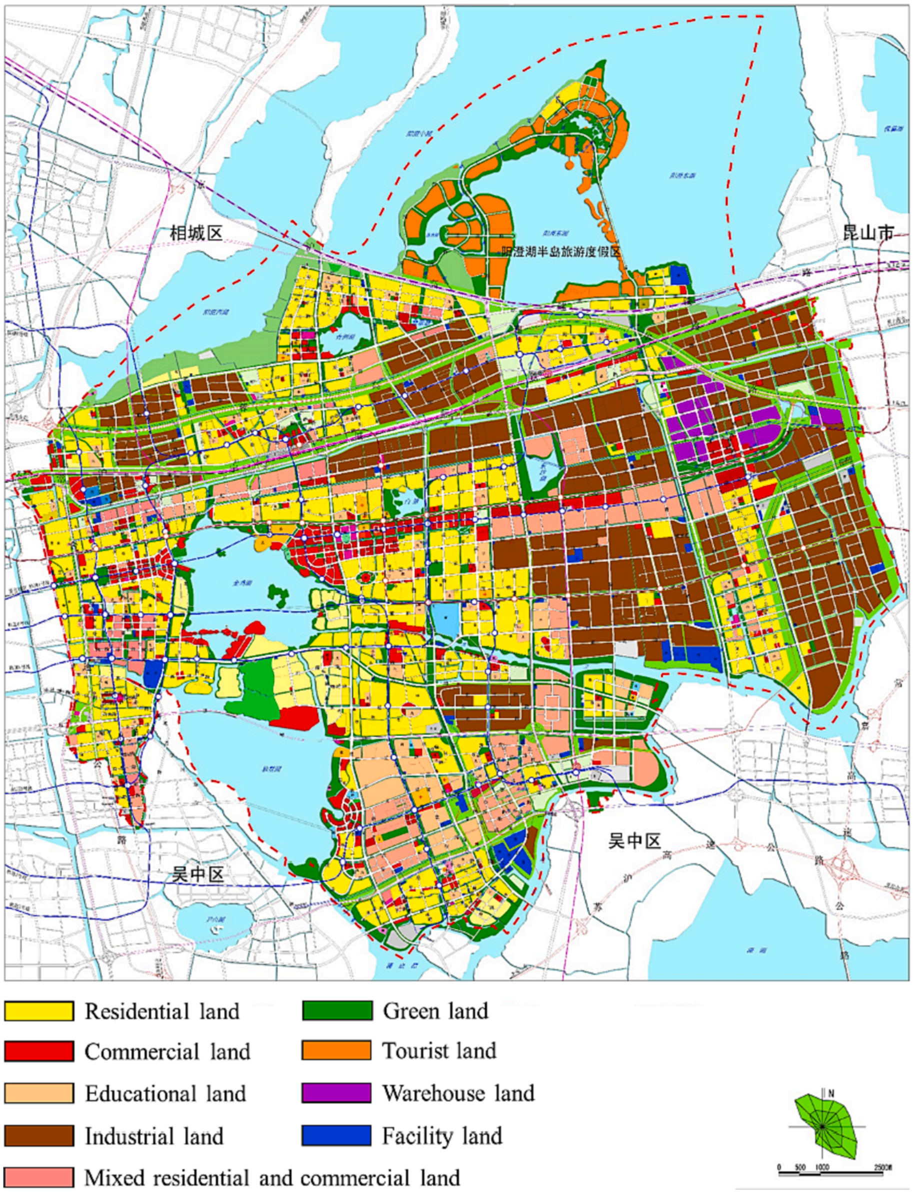

Fig. 3.8. 2012–2030 Land use planning of SIP ().
Source: http://www.sipac.gov.cn/dept/ghjswyh/ghcg/ztgh/201908/t20190822_1049234.htm

Ordinary Least Squares (OLS) is a fundamental model widely applied for curve fitting. It requires the fewest necessary model conditions and finds the best functional match by minimizing the sum of squares of errors. However, very few studies have applied single linear regression to complex accident factor analysis, since the dependent variable in this context is often a discrete value, which makes OLS no longer applicable (Mannering and Bhat, 2014). Therefore, the generalized linear regression model is widely used in accident analysis. Poisson regression requires equalization of the mean and variance, while actual data rarely conform to this assumption. For the over-discrete data, the NB regression model is an effective solution, as shown in equation (7). It adds a disturbance term to represent the discrete effect based on the original Poisson regression.

$$\boldsymbol{E}(\boldsymbol{y}_i|\boldsymbol{x}_i) = \boldsymbol{log\lambda}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\boldsymbol{x}_{i,1} + \cdots + \boldsymbol{\beta}_k\boldsymbol{x}_{i,k} + \boldsymbol{\sigma e}_i \tag{7}$$

Considering the over-dispersion of accident data, we adopted the NB regression model in this study. As a global model, NB assumes that the relationship between the dependent variable and the independent variables is consistent across the entire study space. To capture spatial variations in the relationships between independent and dependent variables, an intuitive solution is to allow the coefficients to vary across the study area (Brunsdon et al., 1996). Fotheringham et al. (2002) first defined the Geograpically Weighted Regression (GWR) model, which introduced spatial weights to describe the spatial relationship (Yue et al., 2017). A GWR model can be expressed as follows. Assuming that there are n observations, for observation $i \in \{1, 2, \cdots, n\}$ at location (u_i, v_i), where $\beta_{i0}(u_i, v_i)$ is the intercept, x_{ij} is the j-th predictor variable, $\beta_j(u_i, v_i)$ is the jth coefficient, ε_i is the error term, and y_i is the response variable,

$$y_i = \beta_{i0}(u_i, v_i) + \sum_{k=1}^{n} \beta_{ik}(u_i, v_i)x_{ik} + \varepsilon_i \tag{8}$$

where y_i represents the number of accidents at sample location i, $\beta_{i0}(u_i, v_i)$ is the constant at location i, (u_i, v_i) represents the coordinate of location i, $\beta_{ik}(u_i, v_i)$ is the coefficient of the k-th independent variable at location i, x_{ik} is the k-th independent variable at sample location i, n is the total number of independent variables, and ε_i is the random error term at sample location i.

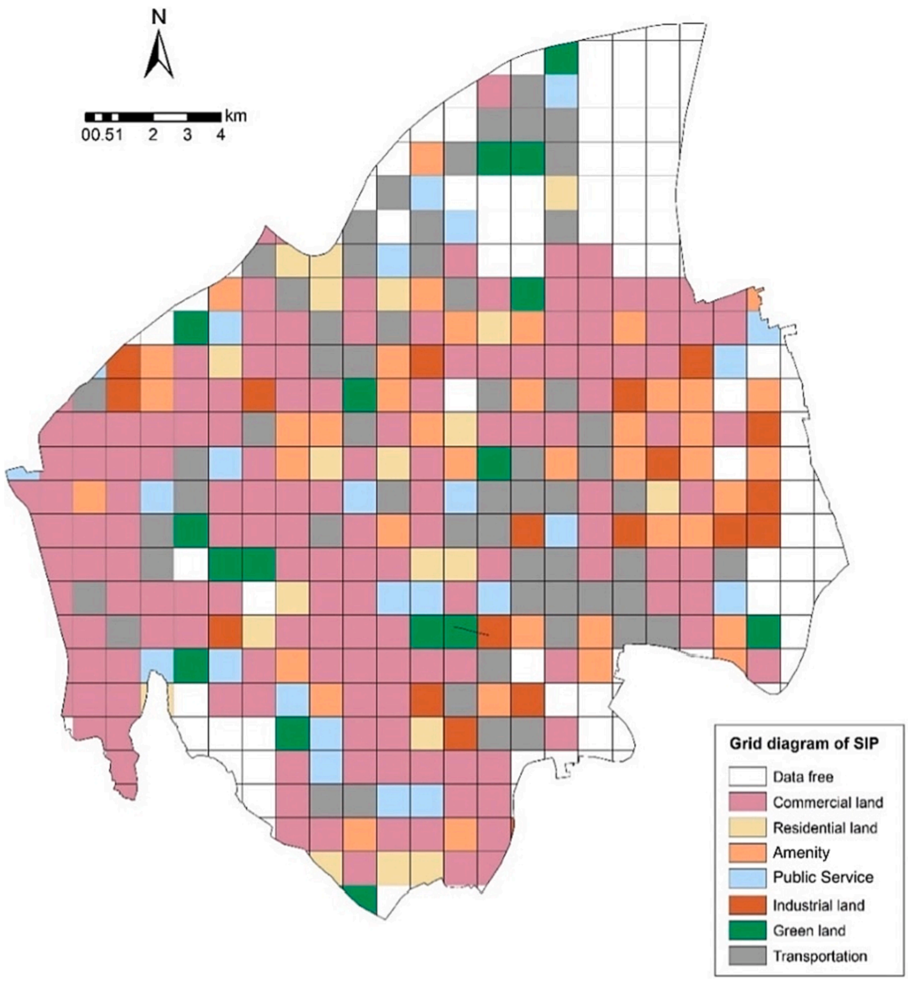

Fig. 3.9. Land use map of SIP based on POI data identification – Grid.

Table 3.1
POI classification.

First level	Second level	Third level
Commercial land	Shopping, hotel and leisure, finance	Supermarkets, shopping centers, restaurants, hotels, cinemas, banks, etc.
Residential land	Residential area, business residential area	Communities, apartments, villa, dormitories, etc.
Amenity	Service shops	Car maintenance, hairdressing, photography, telecommunication offices, travel agencies, etc.
Public Service	Government institutions, medical care, public facilities, schools, etc.	Government agencies, social organizations, emergency centers, schools, etc.
Industrial land	Companies, factories	Companies, factories, science and technology parks, industrial parks, etc.
Green land	Tourist attractions, parks, squares	Scenic spots, zoos, botanical gardens, parks, squares, etc.
Transportation	Road services, road facilities	Bus stops, parking lots, train stations, subway stations, etc.

However, the classical GWR model is only applicable when the data follows the Gaussian distribution. Da Silva and Rodrigues (2014) further extend the traditional GWR model by proposing the Geographically Weighted Negative Binomial Regression (GWNBR) model. It not only addresses spatial dependency and heterogeneity problems for discrete count data, but also takes into account the over-dispersion of the data. The form of the GWNBR model can be expressed as follows:

$$y_i\ NB\left[t_i exp\left(\sum_{k=1}^{n}\beta_{ik}(u_i,v_i)x_{ik}\right),\alpha(u_i,v_i)\right] \tag{9}$$

where y_i is the i-th dependent variable, t_i is an offset variable, (u_i, v_i) are the location of data point i, β_k is the coefficient for the k-th independent variable, and α is the parameter of overdispersion.

Essentially, the GWR model family follows Tobler's First Law of Geography which states that everything is related to everything else, but near things are more related than distant things. This law is applied when calculating the spatial weight matrix, which plays an important role in the estimation of local parameters for each point in the study space. This study used the fixed gaussian kernel function to determine the spatial weight matrix, which can be expressed as follows:

$$W_{ij} = exp\left\{-0.5\left(\frac{d_{ij}}{b}\right)^2\right\} \tag{10}$$

Where d_{ij} represents the Euclidian distance between point i and j, parameter b represents the kernel size. In this study, the Akaike Information Criterion (AIC) is selected for bandwidth optimization because it can trade off the local degrees of freedom and the goodness of fit.

Both the NB and GWNBR models were fitted using the latest version of a set of SAS macros originally developed by Da Silva and Rodriguez (2016), via SAS 9.4 software. At the time this paper was written, it is the only software implementation of the GWNBR model in the research community. The popular MGWR 2.2.1 software does not support the GWR model incorporating NB distribution. The statistical significance of

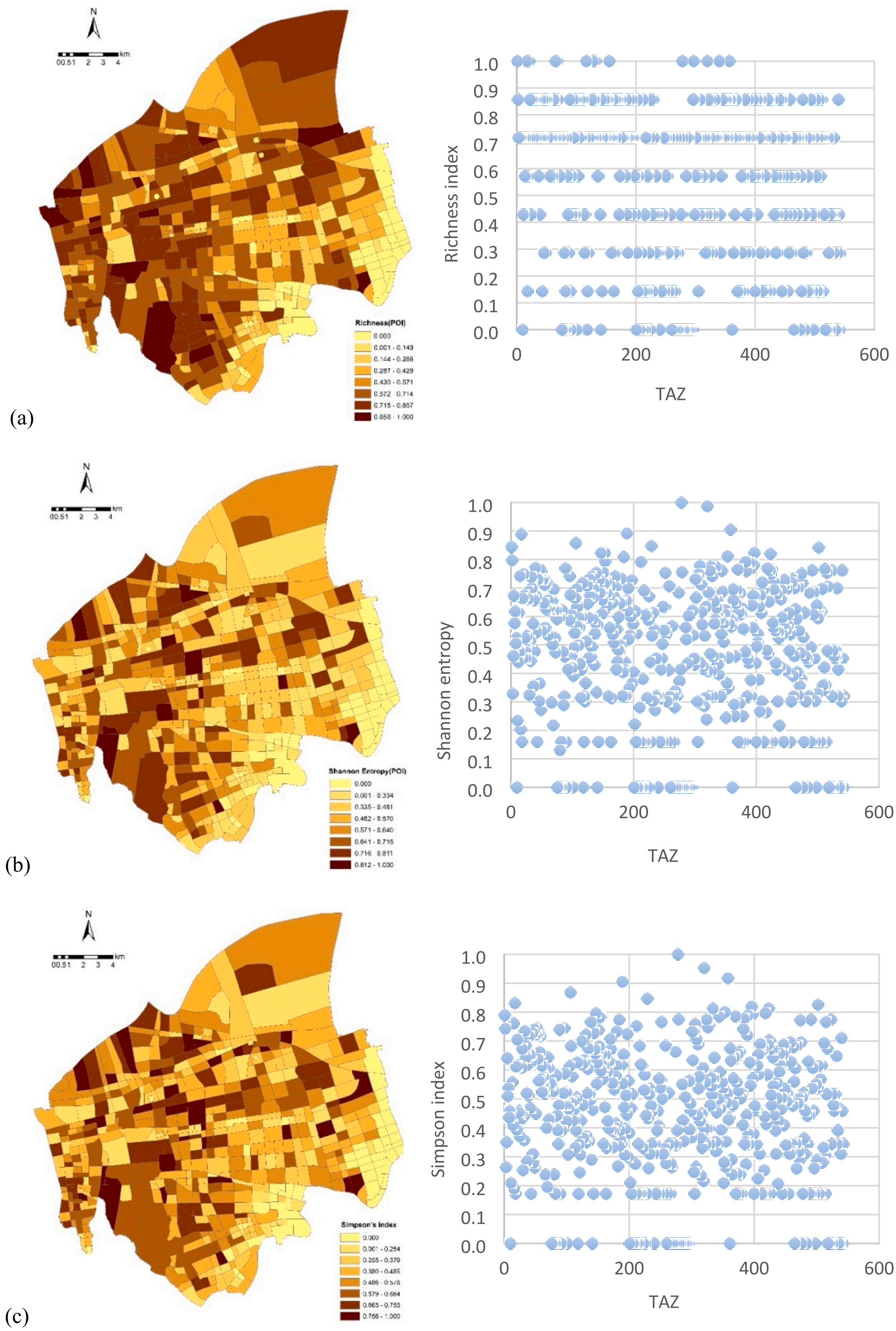

Fig. 3.10. POI-based land use diversity map in SIP (a) Richness index; (b) Shannon Entropy; (c) Simpson index.

the local coefficient estimates was evaluated using tests developed by Da Silva and Fotheringham (2016) which incorporated corrections to avoid false positives in GWR.

4. Results

4.1. Negative binomial regression (Global Model)

According to the NB regression results (Table 4.1), 5 of the 10

Table 3.2
Descriptive statistics of variables.

Variables	min	max	mean	std	Lower quartile	Median	Upper quartile
Accident	1	1636	112.2	193.53	12	38	129.8
Commercial	0.00	2392.41	34.04	148.72	0.00	0.00	31.12
Residential	0.00	180.98	4.98	19.32	0.00	0.00	0.00
Amenity	0.00	508.14	6.58	30.14	0.00	0.00	0.00
Industrial	0.00	222.21	10.8	33.76	0.00	0.00	0.00
Green	0.00	442.02	5.07	31.90	0.00	0.00	0.00
Transportation	0.00	187.75	5.70	21.05	0.00	0.00	0.00
Richness (POI)	0.00	1.00	0.55	0.27	0.29	0.71	0.71
Shannon (POI)	0.00	1.00	0.46	0.22	0.31	0.51	0.66
Simpson (POI)	0.00	1.00	0.43	0.21	0.29	0.46	0.61
Population density	401.8	8218.31	2564.30	2690.61	754.10	871.50	4887.70
Road density	0.00	0.27	0.28	0.05	0.05	0.08	0.11
Restaurant density	0.00	7.77	0.28	0.80	0.00	0.02	0.17

explanatory variables are significant. Among all the land use type indicators, commercial land use, green space, and transportation land use are all significantly associated with traffic accidents with a positive impact, which supports some of the existing studies, but also contradicts findings from some other studies (Ukkusuri et al., 2012; Peera et al., 2019; Kim et al., 2010). The richness index is the only mixed land use index among the three indexes that shows a significant impact (positive) on the number of traffic accidents. The result also demonstrated that population density and road density were positively associated with the number of traffic accidents, with the latter having the strongest impact ($\beta = 2.755$) among all the covariates.

4.2. Geographically weighted negative binomial regression (local model)

Table 4.2 above shows that descriptive statistics of the estimated parameters for the sample points by the GWNBR model, which shows the spatially varying impact that the independent variables have on the number of traffic accidents. It is worth noting that the estimated coefficients of the traditional global model will fall within the range of the corresponding spatial model (Xu and Huang, 2015), indicating that the global model is the result of the average effects of the factors on the dependent variable (Hezaveh et al., 2019). The sign of the median of all the variables in the GWNBR model was found to be consistent with the NB model, thereby proving the above.

Among all the land use type indicators, commercial land use, green land, and transportation land use are seen to have significant impact on the number of traffic accidents locally. Fig. 4.1 below shows the local effects that commercial land use type has on traffic accidents. It can be seen that the impact is only significant in the southern part of SIP and gets, which almost coincides with SIP's Science and Innovation District (also known as the Higher Education Town). This area traditionally has very weak presence of commercial activities (in the form of neighbourhood centres) and large concentrations of higher education and research institutes, universities, and high-tech companies. Commercial types of land use may significantly help to attract pedestrians and traffic volume to these areas, thereby increasing the risk of accidents.

Table 4.1
NB regression analysis results.

Variables	Coef	Std.Err	p-value
Intercept	3.820	0.182	$<0.001^{***}$
Commercial land	0.001	0.0004	0.014^{*}
Residential land	0.00096	0.0034	0.776
Amenity	−0.00097	0.002	0.654
Industrial land	0.00087	0.002	0.6631
Green land	0.0037	0.002	0.072
Transportation	0.0075	0.003	0.015^{*}
Richness (POI)	0.549	0.272	0.043^{*}
Population density	0.000068	0.000025	0.007^{***}
Road density	2.755	1.558	0.077^{*}
Restaurant density	0.027	0.084	0.752

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Fig. 4.2 below shows that green land has positive impact on the number of traffic accidents in the mid-western part of SIP, which covers some of the largest residential communities and the CBD area of the district. This unexpected result is slightly counterintuitive, which would require further study. It could be related to the urban morphology of the area. In such an area with high population density and economic activities, urban green space tends to be very small and scattered throughout the area (i.e., not large enough to reduce/divert traffic). Green space in areas with high density and traffic volume may also decrease people's awareness of safety, thereby increasing the risk of accidents.

Fig. 4.3 below shows that transportation land use only has positive impact on the number of traffic accidents in large areas of the western parts of SIP, and the impact decreases from west to the east. This area houses large residential communities and two of the districts CBD areas (Hudong and Huxi). The west-most area with the strongest impact sits some of the largest and oldest residential communities of SIP, and was among the first to be developed as a new district in the early 1990s. It is also adjacent to Suzhou's old town and is divided by an inner-city and an intra-city expressway, which takes up large amount of space and makes the area less walking-friendly, thereby increasing safety risk. The early development also means that transportation infrastructure is ageing and can become risk factor for road traffic safety (Fig. 4.4).

Among the mixed land use indexes, the Richness Index were found to have significant local impacts in the northwestern part of SIP. Land use types in this area tend to be predominantly leisure (Yangcheng Lake Scenic Area) or industrial, without much variety. It is thought that having more types of POIs may lead to more types of activities in the area, thereby increasing vitality, which could further pose threat on road traffic safety.

Population density is found to have small positive impact on the number of traffic accidents in the central area of SIP, and the impact gets stronger from the northwest to the southeast (Fig. 4.5). The variation of this variable among different TAZs is not large. It can be inferred that in areas of high population density, population density has a positive overall impact on the number of traffic accidents. The impact of population density on traffic accidents is not significant in the narrow eastmost part of the SIP, which could be because other indicators playing a more significant role. Road density plays a significant role on traffic safety in a small area of southwest SIP, where it negatively impacts traffic safety to a great extent (Fig. 4.6).

4.3. Results comparison

Two models, namely Negative Binomial Regression (NB) and Geographically Weighted Negative Binomial Regression (GWNBR) were fitted on the same set of explanatory variables. The NB model was chosen in this study because the accident data was over-discrete and did not meet the Poisson regression hypothesis. Before inputting them into

Table 4.2
Descriptive Statistics of GWNBR model parameter estimates.

Variables	Min	Max	Mean	Q1	Median	Q3
Intercept	3.487	4.222	3.893	3.757	3.891	4.032
Commercial land	0.000387	0.00267	0.00176	0.00127	0.00188	0.00224
Residential land	−0.00182	0.00287	0.00158	0.00136	0.00176	0.00191
Amenity	−0.00471	0.00177	−0.00153	−0.00263	−0.00152	−0.000523
Industrial land	−0.00235	0.00459	0.00128	−0.000095	0.00130	0.00233
Green land	−0.00171	0.00757	0.00436	0.00329	0.00130	0.00564
Transportation	0.000730	0.0171	0.00761	0.00439	0.00741	0.00986
Richness (POI)	0.105	1.348	0.451	0.321	0.424	0.574
Population density	0.000053	0.0000933	0.0000665	0.000061	0.0000649	0.0000721
Road density	1.066	6.992	3.008	1.879	2.846	3.811

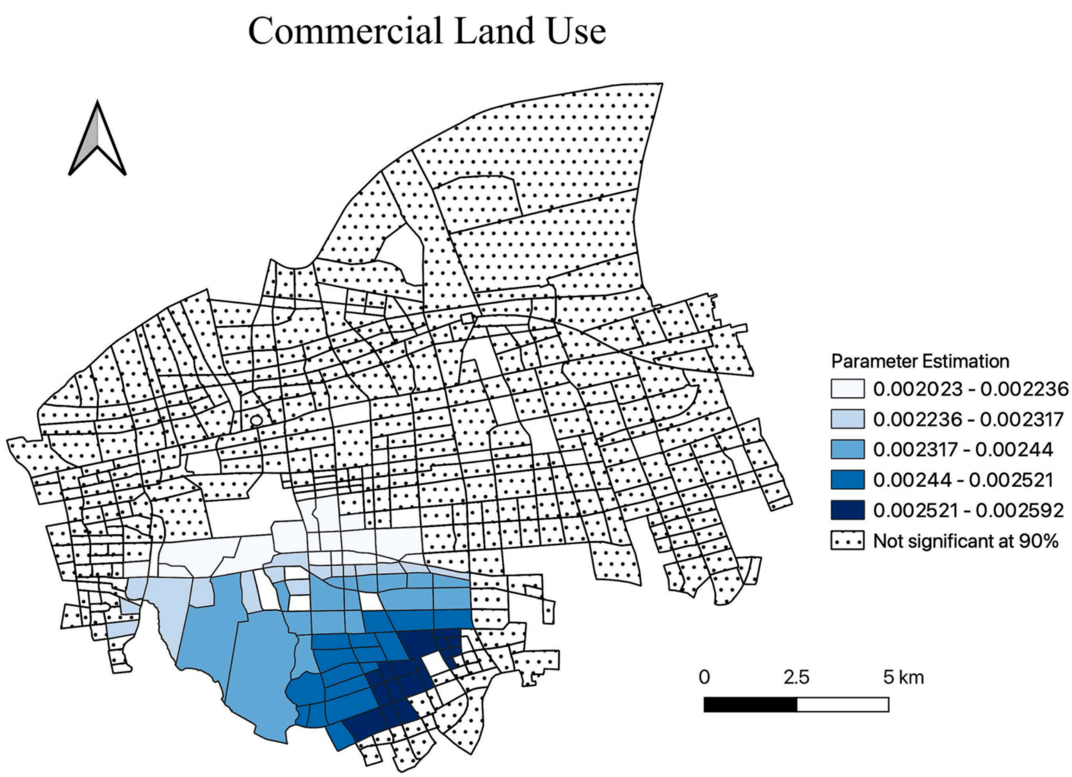

Fig. 4.1. Parameter Estimation for Commercial Land Use.

the model, all variables were standardized. In addition, because the three indicators of land use diversity were significantly correlated, we ultimately selected the Richness index as the only indicator to measure diversity after comparing their different impacts on all the models. The modelling results are presented in Table 4.2. According to the AIC values, it was clear that the GWNBR model is superior to the NB global model.

Table 4.3 shows the comparison of the spatial dependence of the two models using the Moran's I index. It can be seen that the residuals of the NB model are spatially autocorrelated while those of the GWNBR model are not. This provides evidence of the local model tackling spatial heterogeneity and its ability to produced less biased estimates of parameters.

5. Findings and suggestions

To reduce the potential risk of accidents caused by SIP's urban land layout, it is possible to combine the results of our analysis with the city's land use planning (refer to Fig. 3.8).

First of all, SIP could start implementing improvements on some of its ageing transportation infrastructure for them to meet modern day safety standards. Measures can also be done to make areas that are dominated by big transportation infrastructure (e.g., elevated expressway, flyovers) more pedestrian friendly and safely.

Secondly, we suggest improving SIP's public transport system to guide traffic behavior and to reduce the traffic flow dominated by cars so as to reduce the accidents risk associated with mixed land use. According to the analysis, high level of land use mixture significantly increases the risk of accidents. Therefore, while rationally allocating residential, commercial, and industrial land, measures should be taken to improve the urban public transport system and to advocate green travel by residents in order to alleviate traffic congestion and improve road safety.

Third, the scattered distribution of urban parks in SIP increases the concentrated distribution of local residents' leisure and entertainment areas, which not only increases the traffic flow but also centralizes a large number of pedestrian activities. This study shows that scattered green space may pose risk on road traffic safety. However, a more dispersed distribution of green space will make it difficult for residents to meet their leisure needs, thereby increasing the need for residents to travel long distances for recreation while failing to play a role in improving road safety.

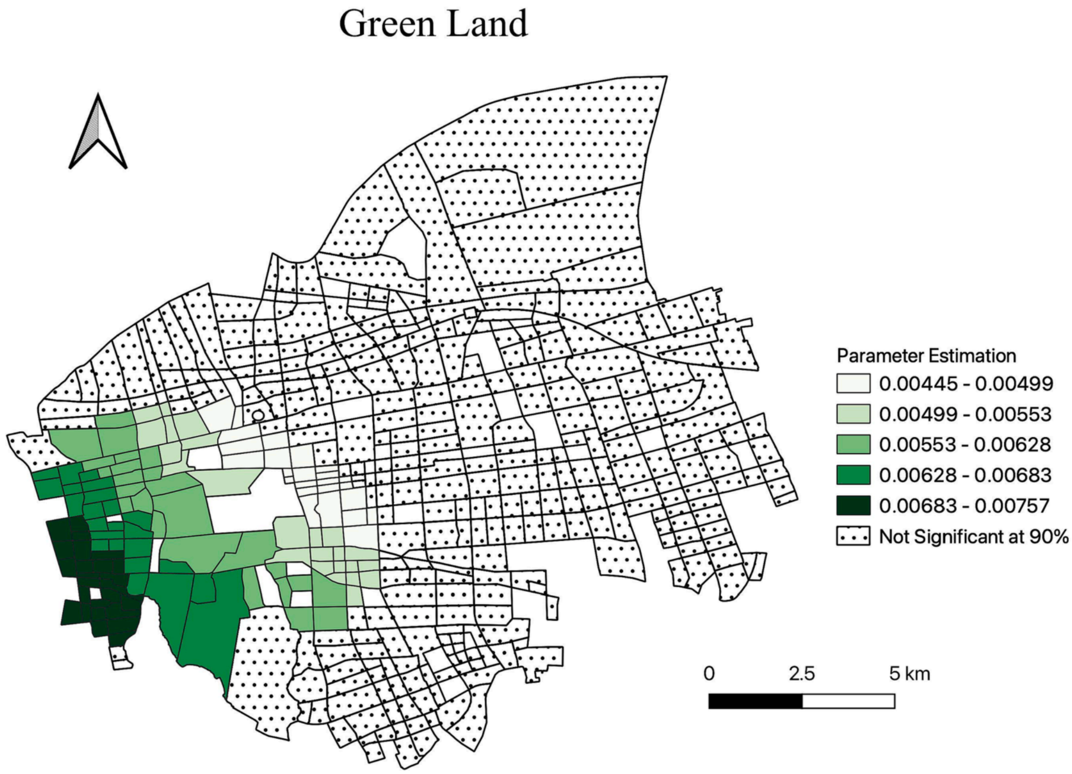

Fig. 42. Parameter Estimation for Green Land.

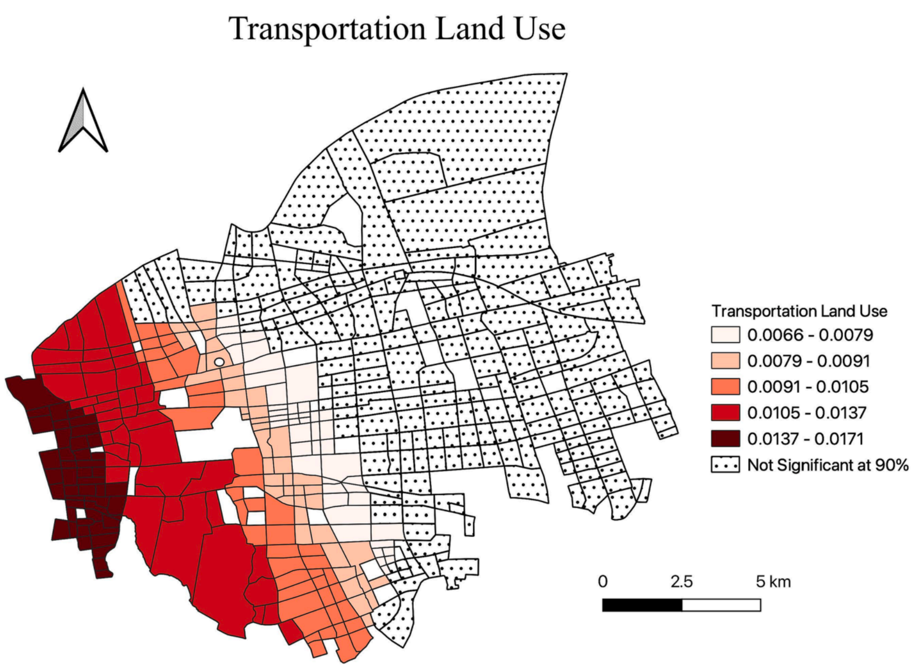

Fig. 4.3. Parameter Estimation for Transportation Land Use.

6. Conclusion

In this study, we analyzed the effects of land use on traffic accidents in SIP based on POI data. To identify the land use characteristics, we first divided the POI data into seven categories and used the TF-IDF algorithm method to calculate the weights of the different POI categories for each TAZ. To explore the effects of POI-based land use on accidents, we utilized NB models for local analysis, while the GWNBR model was employed to introduce a spatial perspective on the effects. Our findings demonstrated that GWNBR model outperforms global regression models, which thus verified our assumption that inherent non-stationarity exists in the estimation of parameters across space.

The results are broadly consistent with existing literature. The findings from the local regression showed that population density and road length are positively correlated with the frequency of traffic accidents. Commercial land use, green land, transportation land use has

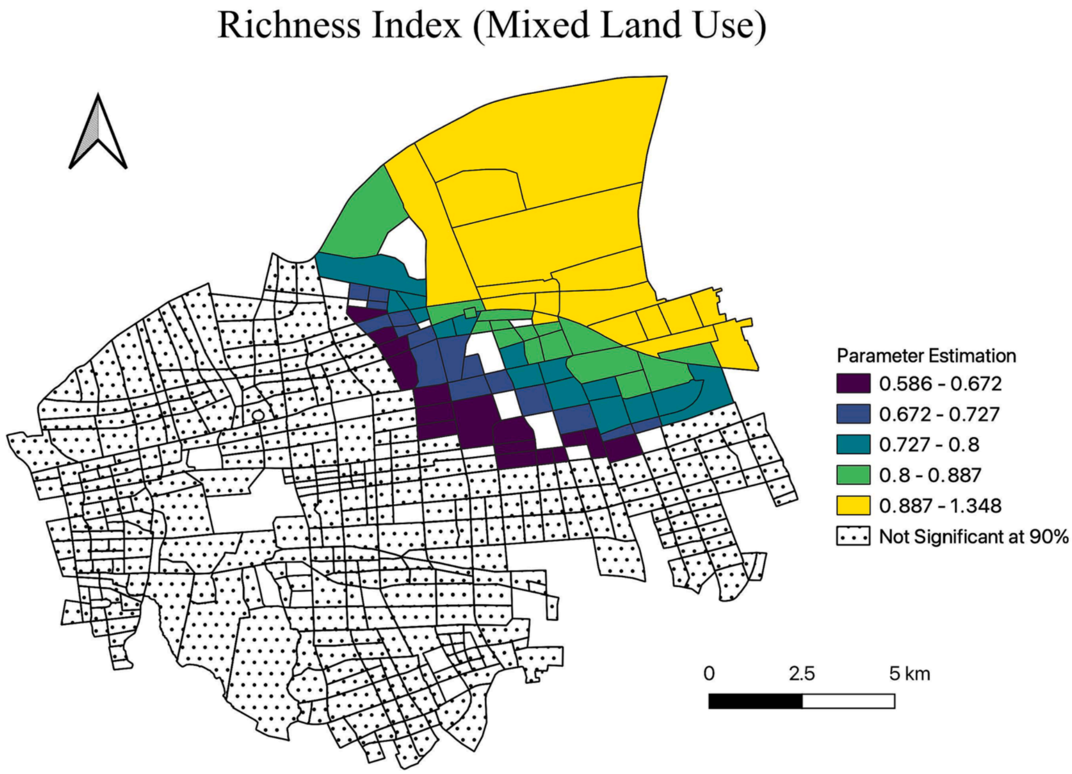

Fig. 4.4. Parameter Estimation for Richness (POI).

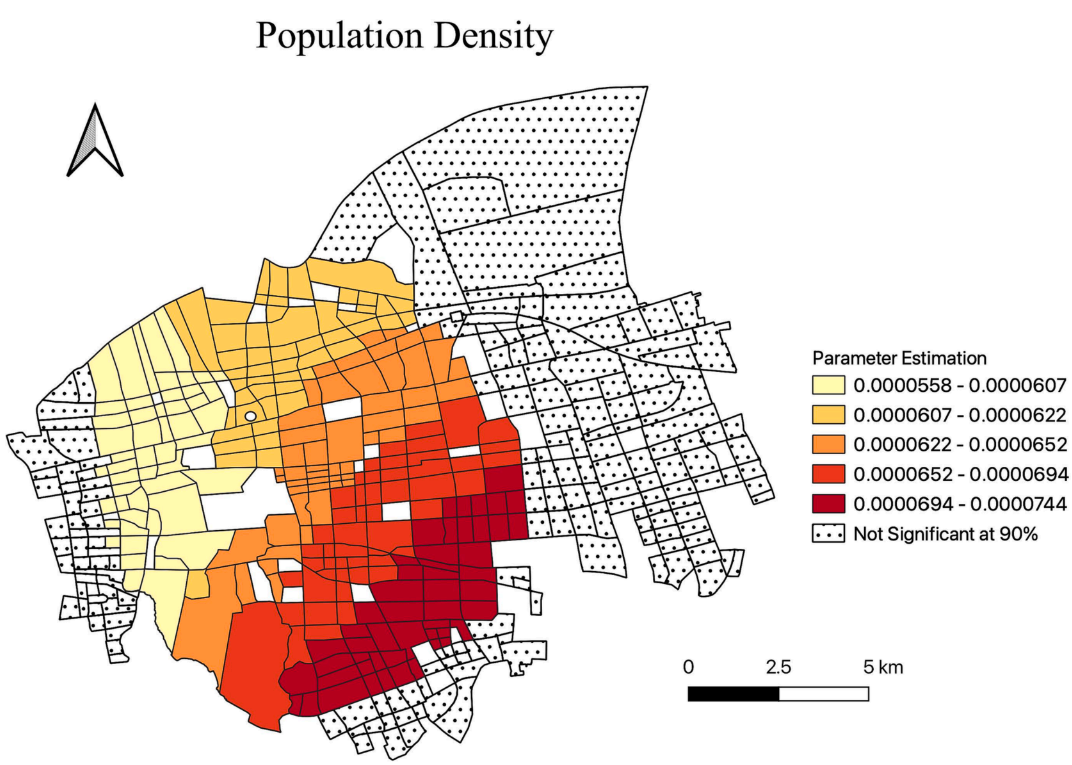

Fig. 4.5. Parameter Estimation for Population Density.

negative impacts on road traffic safety while the other land use types are not significantly associated with accident occurrence. The Richness index which measures the diversity of land use types also proved to have a negative impact on traffic safety. The results from the GWNBR local model provided valuable insight on how the impact of our selected land use and density indicators vary across the entire study space. Some of the indicators showed strong local effects which can be partially explained by morphological and historical reasons. These findings have proved to be valuable in land use and transportation planning. More works needs to be done to verify our assumptions for the rationale

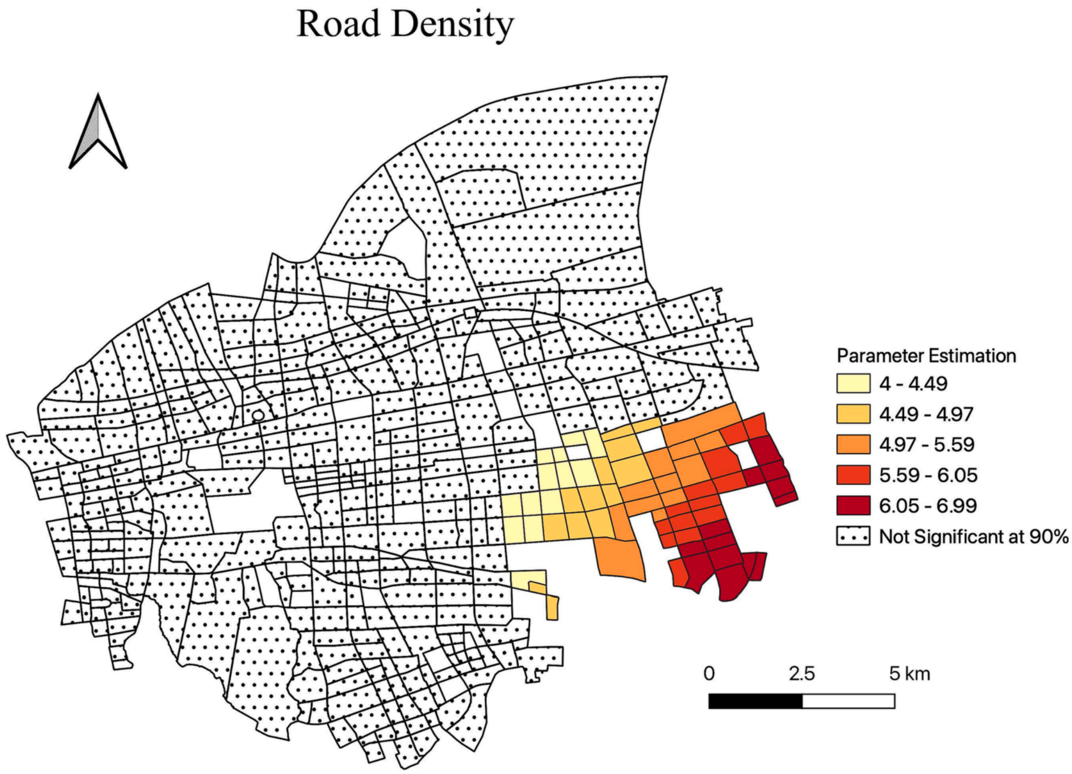

Fig. 4.6. Parameter Estimation for Road Density.

Table 4.3
Comparison of statistics of model fitting.

Model	2 Log-Likelihood	AIC	AICc	Bandwidth	Observations	Dependent variables
NB	−4809.9160	**4833.9**	4834.66	–	434	10
GWNBR	−2392.514	**4826.57**	4828.77	6.04		

Table 4.4
Spatial dependence of residuals of the two models.

Model	Moran's I	p-value
NB	0.039986	<0.001
GWNBR	0.005897	0.488

behind the spatial variations of the impacts.

CRediT authorship contribution statement

Hyungchul Chung: Conceptualization, Methodology, Investigation, Supervision. **Qiaonan Duan:** Data curation, Formal analysis, Investigation, Visualization, Writing – original draft. **Zihao Chen:** Data curation, Formal analysis, Investigation, Project administration, Software, Visualization, Writing – review & editing. **Yueming Yang:** Resources, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Research Institute of Future Cities at Xi'an Jiaotong-Liverpool University and XJTLU Key Programme Special Fund (ref number: KSF-E-43). This paper reflects the views of the authors, who are solely responsible for the facts and the accuracy of the information presented. Xi'an Jiaotong-Liverpool University assumes no liability for the content or uses thereof.

This work was supported by Research Institute of Future Cities at Xi'an Jiaotong-Liverpool University.

References

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accid. Anal. Prev. 38 (3), 618–625. https://doi.org/10.1016/j.aap.2005.12.006.

Anarkooli, A.J., Hosseinpour, M., Kardar, A., 2017. Investigation of factors affecting the injury severity of single-vehicle rollover crashes: A random-effects generalized ordered probit model. Accid. Anal. Prev. 106, 399–410. https://doi.org/10.1016/j.aap.2017.07.008.

Bao, J., Xu, C., Liu, P., Wang, W., 2017. Exploring bikesharing travel patterns and trip purposes using smart card data and online point of interests. Netw. Spat. Econ. 17 (4), 1231–1253. https://doi.org/10.1007/s11067-017-9366-x.

Becker, R.A., Caceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C., 2011. A tale of one city using cdhlar network data for urban planning. IEEE Pervasive Comput. 10 (4), 18–26. https://doi.org/10.1109/MPRV.2011.44.

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically weighted regression: A method for exploring spatial nonstationarity. Geogr. Anal. 28 (4), 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x.

Casares, B.J., Fernández-Aracil, P., Ortuño-Padilla, A., 2019. Built environment and tourism as road safety determinants in Benidorm (Spain). Eur. Plan. Stud. 27 (7), 1314–1328. https://doi.org/10.1080/09654313.2019.1579784.

Cervero, R., Kockelman, K., 1997. Travel demand and the 3Ds: density, diversity, and design. Transp. Res. Part D: Transp. Environ. 2 (3), 199–219. https://doi.org/10.1016/S1361-9209(97)00009-6.

Chen, C., 2017. Improved TFIDF in big news retrieval: An empirical study. Pattern Recogn. Lett. 93, 113–122. https://doi.org/10.1016/j.patrec.2016.11.004.

Chen, H., Chen, H., Liu, Z., Sun, X., Zhou, R., 2020. Analysis of factors affecting the severity of automated vehicle crashes using XGBoost model combining POI data. J. Adv. Transp. 1–12 https://doi.org/10.1155/2020/8881545.

Chen, E., Ye, Z., Wang, C., & Zhang, W. (2019). Discovering the spatio-temporal impacts of built environment on metro ridership using smart card data. *Cities*, 95, pN.PAG-N.PAG. 10.1016/j.cities.2019.05.028.

Chen, P., Zhou, J., 2016. Effects of the built environment on automobile – Involved pedestrian crash frequency and risk. J. Transp. Health 3 (4), 448–456.

Cheng, J., Liu, Y., 2017. Study on cultural distance and cultural adjustment of immigrants from Danjiangkou Reservoir. J. Wuhan Univ. Natur. Sci. Ed. Technol. 30 (4), 135–141. https://doi.org/10.3963/j.issn.1671-6477.2017.04.0021.

Chi, J., Jiao, L., Dong, T., Gu, Y., Ma, Y., 2016. Quantitative identification and visualization of urban functional area based on POI data. J. Geomatics 41 (2), 68–73. https://doi.org/10.14188/j.2095-6045.2016.02.017.

Da Silva, A.R., Rodrigues, T.C.V. (2016), A SAS Macro for Geographically Weighted Negative Binomial Regression (Available from: http://support.sas.com/ resources/papers/proceedings16/8000-2016.pdf).

Da Silva, A.R., Fotheringham, A.S., 2016. The multiple testing issue in geographically weighted regression. Geogr. Anal. 48 (3), 233–247.

Da Silva, A.R., Rodrigues, T.C.V., 2014. Geographically weighted negative binomial regression-incorporating overdispersion. Stat. Comput. 24, 769–783.

Dai, D., Jaworski, D., 2016. Influence of built environment on pedestrian crashes: A network-based GIS analysis. Appl. Geogr. 73, 53–61. https://doi.org/10.1016/j.apgeog.2016.06.005.

Delmelle, E.C., Thill, J.C., Ha, H.H., 2012. Spatial epidemiologic analysis of relative collision risk factors among urban bicyclists and pedestrians. Transportation 39 (2), 433–448.

Ding, C., Chen, P., Jiao, J., 2018. Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: A machine learning approach. Accid. Anal. Prev. 112, 116–126. https://doi.org/10.1016/j.aap.2017.12.026.

Dumbaugh, E., Li, W., 2010. Designing for the safety of pedestrians, cyclists, and motorists in urban environments. J. Am. Plann. Assoc. 77 (1), 69–88. https://doi.org/10.1080/01944363.2011.536101.

Ewing, R., Cervero, R., 2010. Travel and the built environment: A meta-analysis. J. Am. Plann. Assoc. 76 (3), 265–294. https://doi.org/10.1080/01944361003766766.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression. Wiley, New York.

Fotheringham, A.S., Yang, W., Kang, W., 2017. Multiscale geographically weighted regression (MGWR). Ann. Am. Assoc. Geogr. 107 (6), 1247–1265. https://doi.org/10.1080/24694452.2017.1352480.

Gao, X., Zhang, Y., 2018. Study on the health facilities accessibility of the main urban area of Xi'an based on POI data. J. Green Sci. Technol. 6, 212–213.

Handy, S.L., Boarnet, M.G., Ewing, R., Killingsworth, R.E., 2002. How the built environment affects physical activity. Am. J. Prev. Med. 23 (2), 64–73. https://doi.org/10.1016/S0749-3797(02)00475-0.

Hezaveh, A.M., Arvin, R., Cherry, C., 2019. A geographically weighted regression to estimate the comprehensive cost of traffic crashes at a zonal level. Accid. Anal. Prev. 131, 15–24. https://doi.org/10.1016/j.aap.2019.05.028.

Hu, S., Wang, L., 2013. Automated urban land-use classification with remote sensing. Int. J. Remote Sens. 34 (3), 790–803. https://doi.org/10.1080/01431161.2012.714510.

Hu, T., Yang, J., Li, X., Gong, P., 2016. Mapping urban land use by using landsat images and open social data. Int. J. Remote Sens. 8 (2), 151–169. https://doi.org/10.3390/rs8020151.

Huang, Y., Wang, X., Patton, D., 2018. Examining spatial relationships between crashes and the built environment: A geographically weighted regression approach. J. Transp. Geogr. 69, 221–233. https://doi.org/10.1016/j.jtrangeo.2018.04.027.

Jia, R., Khadka, A., Kim, I., 2018. Traffic crash analysis with point-of-interest spatial clustering. Accid. Anal. Prev. 121, 223–230. https://doi.org/10.1016/j.aap.2018.09.018.

Johansson, R., 2009. Vision Zero – Implementing a policy for traffic safety. Saf. Sci. 47 (6), 826–831. https://doi.org/10.1016/j.ssci.2008.10.023.

Kim, K., Brunner, I.M., Yamashita, E.Y., 2006. Influence of land use, population, employment, and economic activity on accidents. Transp. Res. Rec. 1953 (1), 56–64. https://doi.org/10.3141/1953-07.

Kim, K., Pant, P., Yamashita, E., 2010. Accidents and accessibility: Measuring influences of demographic and land use variables in Honolulu, Hawaii. *Transp. Res. Rec.* 2147 (1), 9–17.

Kwan, M.P., 2004. GIS Methods in time – Geographic research: geoeomputation and geovisualization of human activitypatterns. Hum. Geogr. 86 (4), 267–280.

Li, X., Liu, X., Yu, L., 2014. Aggregative model-based classifier ensemble for improving land-use/cover classification of Landsat TM Images. Int. J. Remote Sens. 35 (4), 1481–1495. https://doi.org/10.1080/01431161.2013.878061.

Li, X., Liu, X., Gong, P., 2015. Integrating ensemble-urban cellular automata model with an uncertainty map to improve the performance of a single model. Int. J. Geogr. Inf. Sci. 29 (5), 762–785. https://doi.org/10.1080/13658816.2014.997237.

Li, Z., Wang, W., Pan, L., Bigham, J.M., Ragland, D., 2013. Using geographically weighted poisson regression for county-level crash modeling in California. Saf. Sci. 58, 89–97. https://doi.org/10.1016/j.ssci.2013.04.005.

Liu, Y., Liu, X., Gao, S., Gong, L., et al., 2015. Social sensing: A new approach to understanding our socioeconomic environments. Ann. Assoc. Am. Geogr. 105 (3), 512–530. https://doi.org/10.1080/00045608.2015.1018773.

Louw, E., Bruinsma, F., 2006. From mixed to multiple land use. J. Hous. Built Environ. 21 (1), 1–13. https://doi.org/10.1007/s10901-005-9029-y.

Lyu, C., Wu, X., Liu, Y., Yang, X., Liu, Z., 2020. Exploring multi-scale spatial relationship between built environment and public bicycle ridership: A case study in Nanjing. J. Transp. Land Use 13 (1), 447–467.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22. https://doi.org/10.1016/j.amar.2013.09.001.

McKenzie, G., Janowicz, K., Adams, B., 2014. A weighted multi-attribute method for matching user-generated Points of Interest. Cartogr. Geogr. Inf. Sci. 41 (2), 125–138. https://doi.org/10.1080/15230406.2014.880327.

Osama, A., Sayed, T., 2017. Macro-spatial approach for evaluating the impact of socio-economics, land use, built environment, and road facility on pedestrian safety. Can. J. Civ. Eng. 44 (12), 1036–1044. https://doi.org/10.1139/cjce-2017-0145.

Pan, Y., Chen, S., Li, T., Niu, S., Tang, K., 2019. Exploring spatial variation of the bus stop in fluence zone with multi-source data: a case study in Zhenjiang, China. J. Transp. Geogr. 76, 166–177. https://doi.org/10.1016/j.jtrangeo.2019.03.012.

Pan, Y., Chen, S., Niu, S., Ma, Y., Tang, K., 2020. Investigating the impacts of built environment on traffic states incorporating spatial heterogeneity. *J. Transport Geogr.*, 83, pN.PAG-N.PAG. 10.1016/j.jtrangeo.2020.102663.

Patil, G.R., Pawar, D.S., 2016. Microscopic analysis of traffic behavior at unsignalized intersections in developing world. Transport. Lett. 8 (3), 158–166. https://doi.org/10.1179/1942787515Y.0000000019.

Peera, K.M., Shekhawat, R.S., Prasad, C., 2019. Traffic analysis zone level road traffic accident prediction models based on land use characteristics. Int. J. Traffic Transport Eng. 9 (4), 376–386. https://doi.org/10.7708/ijtte.2019.9(4).03.

Pulugurtha, S.S., Duddu, V.R., Kotagiri, Y., 2013. Traffic analysis zone level crash estimation models based on land use characteristics. Accid. Anal. Prev. 50 (1), 678–687. https://doi.org/10.1016/j.aap.2012.06.016.

Qian, X., Ukkusuri, S.V., 2015. Spatial variation of the urban taxi ridership using GPS data'. Appl. Geogr. 59, 31–42. https://doi.org/10.1016/j.apgeog.2015.02.011.

Saha, D., Dumbaugh, E., Merlin, L.A., 2020. A conceptual framework to understand the role of built environment on traffic safety. J. Saf. Res. 75, 41–50. https://doi.org/10.1016/j.jsr.2020.07.004.

Sun, L.S., Wang, S.W., Yao, L.Y., Rong, J., Ma, J.M., 2016. Estimation of transit ridership based on spatial analysis and precise land use data. Transport. Lett. 8 (3), 140–147. https://doi.org/10.1179/1942787515Y.0000000017.

Tercan, E., Beşdok, E., Tapkin, S., 2020. Heuristic modelling of traffic accident characteristics. Transport. Lett. 1–9 https://doi.org/10.1080/19427867.2020.1734273.

Tu, M., Chen, Z., 2020. Spatial layout and location characteristics of shanghai logistics industry based on POI data. Logistics Technol. 39 (8), 62–67. https://doi.org/10.3969/j.issn.1005-152X.2020.08.013.

Ukkusuri, S., Hasan, S., Aziz, H., 2011. Random parameter model used to explain effects of built-environment characteristics on pedestrian crash frequency. Transp. Res. Rec. 2237, 98–106. https://doi.org/10.3141/2237-11.

Ukkusuri, S., Miranda-Moreno, L.F., Ramadurai, G., Isa-Tavarez, J., 2012. The role of built environment on pedestrian crash frequency. Saf. Sci. 50 (4), 1141–1151. https://doi.org/10.1016/j.ssci.2011.09.012.

Wier, M., Weintraub, J., Humphreys, E.H., Seto, E., Bhatia, R., 2009. An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning Accid. Accid. Anal. Prev. 41 (1), 137–145. https://doi.org/10.1016/j.aap.2008.10.001.

Wu, S., Silvánhyphen, C.J., Wang, L., 2007. Per-field urban land use classification based on tax parcel boundaries. Int. J. Remote Sens. 28 (12), 2777–2801. https://doi.org/10.1080/01431160600981541.

Wu, S., Qiu, X., Usery, E.L., Wang, L., 2009. Using geometrical, textural, and contextual information of land parcels for classification of detailed urban land use. Ann. Assoc. Am. Geogr. 99 (1), 76–98.

Xia, M., Cao, J., Zhou, W., 2014. Analysis on relation between meteorological condition and road traffic accident in Nanjing. J. Meteorol. Sci. 34 (3), 86–90. https://doi.org/10.3969/2014jms.0002.

Xie, B., Nan, X., 2018. The influence of urban land-use on traffic crashes-A case of Wuhan. In: Proceedings of the 2018 China Urban Planning Annual Conference, pp. 1–18.

Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: random parameter versus geographically weighting. Accid. Anal. Prev. 75, 16–25.

Xue, C., Zheng, X., Zhang, B., Yuan, Z., 2015. Evolution of a multidimensional architectural landscape under urban regeneration: a case study of Jinan, China. Ecol. Ind. 55, 12–22. https://doi.org/10.1016/j.ecolind.2015.02.036.

Yang, Y., Chung, H., Kim, J.S., 2021. Local or Neighborhood? Examining the relationship between traffic accidents and land use using a gradient boosting machine learning method: The Case of Suzhou Industrial Park, China. *J. Adv. Transport.*, vol. 2021, Article ID 8246575, 30 pages. https://doi.org/10.1155/2021/8246575.

Yang, X., Wu, S., Zhang, Q., 2018. Analysis of the relevance of spatial distribution between urban parks and transportation infrastructure in Beijing based on POI Data. J. Landscape Res. 10 (3), 27–34. https://doi.org/10.16785/j.issn1943-989x.2018.3.007.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., Mai, K., 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. Int. J. Geogr. Inf. Sci. 31 (4), 825–848. https://doi.org/10.1080/13658816.2016.1244608.

Yue, Y., Zhuang, Y., Yeh, A.G.O., Xie, J., Ma, C., Li, Q., 2017. Measurements of POI-based mixed use and their relationships with neighborhood vibrancy. Int. J. Geogr. Inf. Sci. 31 (4), 658–675. https://doi.org/10.1080/13658816.2016.1220561.

Zhang, Y., Li, Q., Huang, H., Wu, W., Du, X., Wang, H., 2017. The combined use of remote sensing and social sensing data in fine-grained urban land use mapping: A case study in Beijing, China. *Remote Sens. (Basel)* 9 (9), 865–888. https://doi.org/10.3390/rs9090865.

Zhao, J., Li, H., Deng, S., Jiang, J., 2019. Identifying urban functional regions based on POI data and spatial analysis of main transit hubs. Geoma. Spatial Inf. Technol. 42 (12), 38–42.

Zheng, Y., Xie, B., Nan, X., An, Z., 2020. The impact factor and mechanism of urban land use on traffic accidents: a case study of Wuhan City. Modern Urban Res. 2, 42–49. https://doi.org/10.3969/j.issn.1009-6000.2020.02.006.

Further reading

Oshan, T.M., Li, Z., Kang, W., Wolf, L.J., Fotheringham, A.S., 2019. mgwr: A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. ISPRS Int. J. Geo Inf. 8 (6), 269.

Salyon, G., Buckley, C., 1988. Term-Weighting approaches in automatic text retrieval. Inf. Process. Manag. 24 (5), 513–523.

World Health Organization. (2021). *Road traffic injuries*. Available at: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (Accessed: 23 June 2021).